UNDERSTANDING AUTOMATION HANDOFF IMPACTS ON WORKLOAD AND TRUST WHEN MITIGATED BY RELIABILITY DISPLAYS

A Thesis
Presented to
The Academic Faculty

by

Brittany E. Noah

In Partial Fulfillment
of the Requirements for the Degree
Masters of Science in Psychology in the
School of Psychology

Georgia Institute of Technology
August 2018

UNDERSTANDING AUTOMATION HANDOFF IMPACTS ON WORKLOAD AND TRUST WHEN MITIGATED BY RELIABILITY DISPLAYS

Approved by:

Dr. Bruce N. Walker, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Jamie C. Gorman
School of Psychology
Georgia Institute of Technology

Dr. Ayanna M. Howard
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: April 27, 2018

ACKNOWLEDGEMENTS

To my lab mates, thank you for your endless support and listening to me ramble about this for the last year. Thank you to my family and friends for your unwavering support and encouragement throughout this process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

Current commercial vehicles are beginning to include automated features such as adaptive cruise control and automated lane keeping. This is a first step towards full vehicle automation which is predicted to be possible within the next five years. As automated features are integrated into vehicles, the driver must know how to properly interact with and trust these systems. A key element of drivers interacting and relying on these systems is the handover of control between the vehicle and driver. This handover, occurring during times of automation error, will be a critical point of high workload for drivers when driving a partially or fully automated vehicle. If the driver is aware of the system's performance and can appropriately calibrate his or her trust, then these instances of handover may become less stressful and easier to complete successfully. This study explored the driving performance, trust, visual scanning behaviors, perceived workload, and objective workload for handover scenarios. There were four between-subjects display conditions: (1) no display; and reliability displays using (2) quantitative information (percentage of reliability); (3) qualitative information (direct representation of a number); and (4) representational information (abstract representation of a number). Participants completed two drives. The first drive aided in familiarization with the automated lane keeping system. In the second drive, the handover drive, participants experienced an automation failure resulting in transition of control from automated to manual. Results from this study showed that there was a difference in subjective experience between the baseline and handover drive due to experiencing an automation failure. Participants in the no display condition

were more affected by the automation failure, greatly decreasing their overall trust in the automated lane keeping system. Participants with reliability displays were able to appropriately calibrate their trust to system performance and were less impacted by the automation failure, experiencing a slight, statistically insignificant, decrease in trust. These findings will impact the implementation and design of automation reliability displays and shows that drivers with reliability displays are less impacted by automation failure than those without reliability displays.

CHAPTER 1. INTRODUCTION

Automation is becoming more pervasive in our society. Automated vehicles are one instantiation of increased automation transforming the way humans and technologies interact. First implemented in airplanes to reduce pilot workload, automated safety systems and higher levels of automation are also being implemented into commercial vehicles. These systems are aimed at increasing the safety of drivers, yet are not always designed to give drivers feedback on their reliability.

1.1 Automation

Automation is defined as a machine or system doing a task that a human once performed (Parasuraman & Riley, 1997). There are different levels of automation that range from high to low depending on the human and system task requirements (Endsley & Kaber, 1999). For example, at high levels of automation, there is very little engagement of the human with the system carrying out most tasks and making decisions; alternatively, at low levels of automation, the human extensively engages with the system, making decisions and controlling the system through inputs (Endsley & Kaber, 1999).

1.1.1 Automated driving

The Society of Automotive Engineers (SAE) International has defined levels of automation adopted by the United States Department of Transportation for commercial vehicles. These levels range from 0 to 5; where 0 is a vehicle with little to no automation (i.e. automatic transmission, power windows, cruise control) and 5 is a fully autonomous vehicle that can drive entirely on its own once a destination is input by the operator. For

more detailed information on the SAE automation taxonomy, see APPENDIX K. SAE LEVELS OF AUTOMATION. Current commercially available vehicles typically fall within automation levels 1 to 3; however, there are some vehicles such as Tesla's Model S, that have higher levels of automation. This study focuses on Level 2 automation, "Partial Automation: The driving mode-specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the human driver performs all remaining aspects of the dynamic driving task" (SAE J3016).

At SAE's Levels 2 and 3, there is an exchange of information and control of the vehicle between driver and the vehicle. Researchers have categorized these interactions into a taxonomy for better understanding vehicle and driver information requirements and time to react (Mccall, McGee, Meschtscherjakov, Louveton, & Engel, 2016). This research focuses on non-scheduled system initiated handovers. In this scenario, the vehicle determines that it is incapable of continuing to control the vehicle due to something within the driving environment. While these situations may become emergent due to the nature of changing roadway conditions or poor driver reaction time, they are initiated as non-emergent take over requests (Mccall et al., 2016).

As automation becomes increasingly available in vehicles, the driving task is becoming more supervisory in nature. Supervisory control tasks inherently require less action from operators leaving them to scan the operating environment periodically to ensure that all systems are performing at a satisfactory level (Sheridan, 2012). Over long periods of time, these tasks often lead to decrements in performance because operators

decrease vigilance with highly reliable automation in low workload task environments (Sheridan, 2012).

This low workload, vigilance task environment can also lead to lower levels of situation awareness. Situation awareness (SA) is a mental model of a task environment that allows operators to predict future system states (Durso & Gronlund, 1999). Further discussion of SA can be found in section 1.3. This is imperative in the dynamic driving environment as the risks of low SA are accidents, injuries and even fatalities.

1.1.2 Automation reliability

Reliability of automation is defined as performance of the automated system as measured by number of errors made by the system during a given amount of time. Under high levels of automation reliability (corresponding to fewer automation errors), intermittent failures may not have an impact on trust in the system unless the failures continue to occur over time, manifesting themselves at critical points of operation (Parasuraman, 1997; Stanton & Marsden, 1996). In aviation, there are many examples of failed warning systems and over reliance on automated features of flight that when undetected lead to accidents (Stanton & Marsden, 1996). Error detection has been shown to be an important factor in overall human-automation performance for supervisory control tasks (Sheridan, 2012). As an increased number of cars with automated features are being developed, displays will become critical to ensuring drivers understand how to use the technology in their vehicle as well as know when to intervene if it fails.

1.1.3 Automation failures

Automation failures, in tasks perceived as easy by the operator, have been shown to decrease trust and reliance in automated systems (Madhavan, Wiegmann, & Lacson, 2006). As driving requires little formal training to acquire a license, it is likely that many operators will perceive the tasks controlled by the automation to be easy, in turn magnifying the impact of the automation failures.

1.1.4 Handover in automated driving

Changing who or what is in control of an automated system is defined as handover. This change in control can take many forms resulting from automation errors, situation awareness of the automation, and driver preference (Mccall et al., 2016). A taxonomy of handover in automated driving discusses five different types of handover that occur in the transition of control from car to driver (Mccall et al., 2016). First, *scheduled* handover, refers to the vehicle acknowledging that it is going to enter an area that it cannot drive autonomously in. This type of handover allows for advanced alerting of the driver since these areas are identified at the route-planning phase. *Non-scheduled system initiated* handovers occur when the vehicle must relinquish control to the driver unexpectedly. This could be due to a change in conditions on the roadway that were unexpected by the vehicle. These handovers are forced because of the abilities of the automated system and therefore, leave little time to notify drivers in advance. *Non-scheduled user initiated* handovers are those in which the user determines that they would like to take control of the vehicle. This could be due to their desire to control the vehicle, a stop at the grocery store, or any number of things. These handovers are forced by drivers and therefore alerting them to the change in control is unnecessary. *Non-scheduled user initiated emergency* situations occur when the driver becomes aware of a potential risk in the driving environment. *Non-scheduled*

system initiated emergency handovers are the result of errors within the system entirely. These handovers are not due to anything in the external roadway environment and, therefore, cannot continue to operate safely. In these situations, it is assumed that there will be a safe shut off mode, if the system is unable to notify the driver of the failure, such as pulling over to the shoulder and stopping (Mccall et al., 2016). These five handover scenarios describe the ways that the driver or the system can initiate handover of control from automated mode to manual mode.

This study focuses on the non-scheduled system initiated handover context. This context can be further characterized as one in which the driver and the vehicle manufacturer share responsibility for situation awareness of the driver (Mccall et al., 2016). The handover's shared responsibility for SA makes it ideal for studying the impacts of displays on handover transitions.

To improve handover transitions, many researchers have studied how much time drivers need to resume vehicle control. In a review of 16 studies by Eriksson and Stanton (2017), mean take over reaction time was calculated to be 2.96 +/- 1.96s. While this finding gives a good starting point for when to alert drivers prior to an automation failure to ensure safe resumption of vehicle control, the exact driving context and type of automation will impact time required for control resumption. Additionally, there is little known about the time required to return to baseline levels of workload after a handover is completed. This handover recovery time is also essential to understand as additional tasks such as complying with navigation instructions or conversing with passengers may be difficult during this time of high workload.

1.2 Cognitive workload

Cognitive workload is the amount of mental effort required to complete a task. This can be influenced by many factors based on the task itself, the operator's behavior, the operator's performance, and the operator's perception of the task and their own performance (Hart & Staveland, 1988). Complex, dynamic tasks, such as driving, are often high workload. Alternatively, less complex, less dynamic tasks often result in lower workload.

1.2.1 Characteristics of supervisory control tasks

The nature of automation, completing tasks once performed by a human operator or user, suggests a reduction in cognitive workload. This is due to a shift in the role of the operator from direct control to supervisory control of systems. Sheridan (2012) describes supervisory control as occurring in five stages: plan, teach, monitor automation, intervene, and learn. The *planning* stage describes the phase during which operators must plan the task to be undertaken by the automation. The *teaching* phase addresses the operator's task of communicating commands to the automation to complete the planned task. *Monitoring automation* is when operators must maintain awareness of the automation's actions, estimate future states of the automation based on past and present states, and evaluate the state of the automation to ensure that there are no failures or errors in the system. The *intervention* phase occurs if a failure or error is detected in the automation during the monitoring phase. This intervention behavior depends on whether an error or a failure is detected. For error states, troubleshooting of the error should occur. In failure modes, resumption of manual control if possible or abortion of the automated process should occur.

Lastly, the *learning* phase should include record keeping and analysis of the automated process as a whole (Sheridan, 2012).

The level at which each of these phases of automation supervision occurs is dependent on the level of automation in the system. For example, the SAE level 2 automation, as present in this study, does not require the operator to plan or teach the system. The supervisory control task will exist in the later stages only: monitoring automation, intervening, and learning. The primary phase of the supervisory control task for participants will be monitoring automation. This will require scanning of the driving environment for failures or errors. This type of task requires a high degree of vigilance to accurately detect failures and errors of the automation in the driving environment. Prior research has shown, when failures do occur, the workload of the operator will increase significantly causing cognitive workload to be even higher than if in a manual control state (Sheridan, 2012). This is due to the potential shift from little to no attention on the automated portion of the task, to sudden need to gain insight about the environment quickly and accurately and then determine the proper intervention (Sheridan, 2012). This sudden transition will lead to a very rapid rise in workload (Sheridan, 2012).

1.2.2 Cognitive workload measurement

There are many measures of workload that range from subjective to physiologically based. This study will take advantage of subjective and physiological measures to collect real-time workload data as well as perceived (subjective) workload data. The subjective measure to be used is NASA-TLX. This measure was developed by Hart and Staveland (1988) to quantify the subjective experience of workload across a variety of environments.

This measure is comprised of six workload related factors: mental demand, physical demand, temporal demand, performance, effort, and frustration level (Hart & Staveland, 1988). These six items are given ratings by the participant to determine how much of each resource was required for the task completed. In addition, participants are also presented with all possible pairwise comparisons of the six sources of workload and asked to choose which source had more effect on their overall workload. This choice task is used in combination with the rating scale to determine overall workload (Hart & Staveland, 1988).

In addition to subjective measures, physiological measures of workload have also been proposed. These physiological measures have been proposed as dynamic measures of workload. Dynamic measurements improve the timescale of measurement from a single instance after the task has been completed, as is done in subjective measures like NASA-TLX, to continuous measurement throughout the task, which captures changes throughout task completion. Heart rate has been found to increase linearly with cognitive workload in the context of driving (Mehler, Reimer, Coughlin, & Dusek, 2009). Pupil size has also been proposed as a measure of workload in the driving environment (Palinko, Kun, Shyrokov, & Heeman, 2010; Recarte & Nunes, 2003). As task difficulty increases, researchers have shown that pupil size also increases (Granholm, Asarnow, Sarkin, & Dykes, 1996; Iqbal, Zheng, & Bailey, 2004). This measure is reliable but has a ceiling effect at the peak of cognitive load (Iqbal et al., 2004). One area of concern when using pupil size as a measure of workload is pupillary light reflex, which occurs when light levels change throughout a measurement period (Kun, Palinko, & Razumenic, 2012). As participants will only be looking at the driving simulation while pupil size is recorded, the luminance will be constant throughout a measurement period.

1.3 Situation awareness

Situation awareness (SA) is defined as, "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988, pg. 792). This construct has three phases: perception, comprehension, and projection. The perception phase consists of taking in environmental information through the senses and attaching meaning to it the stimuli. The comprehension phase is understanding the properties of the environment based on the perceptions of the stimuli developed in the first phase. Finally, the projection phase builds upon the operator's understanding of the environment to predict future system states (Endsley, 1995).

While all three phases of SA are important for safely operating a semi-automated vehicle, perhaps the most critical is projection. In cases of handover, the ability to anticipate future system states could improve driving performance and safety throughout the handover task. The displays that will be tested in this study were designed to increase SA of the automation's performance throughout the drive by using trend displays (Endsley, 2011). These displays give drivers information about the system's performance throughout the drive so that they can become aware of how it may increase and/or decrease over time.

1.3.1 SA and automation

Automation's effect on SA has been investigated in a number of environments and has been found to improve awareness toward future states through reduction in workload (Billings, 1991; Vortac, Edwards, Fuller, & Manning, 1995). In the Air Traffic Control (ATC) environment, automation was found to improve prediction of future states due to

reduced workload and increased mental resources available for predictive activities (Vortac et al., 1995). There is also evidence indicating that there is a reduced understanding of the environment due to increased automation (Sarter & Woods, 1995). This evidence aligns with Parasuraman and colleagues' theory of over reliance on automation, due to the supervisory nature of highly automated tasks, leading to complacency (Parasuraman, Molloy, & Singh, 1993). Further research suggests that improved displays could reduce the deficit in SA caused by automation (Kieras & Meyer, 1995). From these findings, Durso and Gronlund suggest three characteristics of good automation in terms of maintaining SA: decreased overall workload with some task engagement for improving memory of important tasks, operator awareness of the system's mode, and information tracking for reduced SA due to automation (Durso & Gronlund, 1999).

1.3.2 SA measurement

Measurement of SA can be done through three primary methods: subjective, query-based and implicit performance measures (Durso & Gronlund, 1999). Subjective methods of SA are self-reported levels of situation awareness while completing a task or set of tasks (Durso & Gronlund, 1999). Situation Awareness Rating Technique (SART), is one example of this technique that collects respondents' levels of situation awareness through a set of Likert scale items (Taylor, 1990). These subjective ratings allow for insight into perceived SA of participants but are reliant upon memory of the task and can only be administered at certain time points rather than throughout a task.

Query based methods such as the Situation Present Assessment Method (SPAM) and Situation Awareness Global Assessment Technique (SAGAT) are often used in the

evaluation of user interfaces such as those evaluated in this study (Durso et al., 2015; Endsley, 1988). While the query technique can pinpoint specific areas of the display or levels of SA, they can lead to priming participants for potential events in dynamic environments, especially when measuring Level 3 SA.

Performance measures do have significant advantages (Endsley, 1995). These measures allow for objective, although indirect, measurement of SA throughout a task. While global performance measures suffer from insensitivity, imbedded task measures allow for higher sensitivity through subtask performance measurement (Endsley, 1995). These imbedded task measures allow for pinpointing SA surrounding specific subtasks that may be important for safety, overall task performance, or may be impacted by a new strategy or display (Endsley, 1995). Because of their specificity, imbedded task measures are unlikely to truly measure overall SA in complex task environments, however, their dynamic nature allowing for measurement over time, improves measurement fidelity and sensitivity in complex environments over subjective measures (Endsley, 1995). These measures also have the advantage over query methods in that they are nonintrusive and therefore will not prime operators to behave any differently than they would outside of the laboratory environment (Endsley, 1995).

Eye movements have traditionally been used to measure workload (Bartels & Marshall, 2012; Dehais, Causse, & Tremblay, 2011; Ellis, 2009; Ikuma, Harvey, Taylor, & Handal, 2014; Maier, Baltsen, Christoffersen, & Strrle, 2014; Palinko et al., 2010). Research has shown that decreased spread in gaze fixation is due to high workload (Recarte & Nunes, 2003). This finding indicates that users are taking in less information from their environment due to high workload, thereby reducing their awareness of their environment

(SA). Wickens and colleagues proposed an Attention-Situation Awareness model to link attention allocation (as measured by eye movements) to SA (Wickens, 1996). This model, based on the Salience, Effort, Expectancy, and Value (SEEV) Model for modelling visual attention, was developed specifically to model human errors in SA in which attention is not properly allocated (Wickens et al., 2007). Errors at Level 1 SA (perception) were found to be responsible for a majority of aviation accidents (Jones & Endsley, 1996). Errors at Level 1 SA often occur because of attentional tunnelling (or narrowing of gaze fixation) (Sarter & Woods, 2000). The ability to appropriately allocate attention while supervising automation has been shown to be incredibly important (Parasuraman, Sheridan, & Wickens, 2000; Sarter & Woods, 1995). Therefore, using gaze fixation spread as a measure of attention allocation could indirectly serve as a dynamic measure of situation awareness.

1.4 Trust in automation

Though originally an interpersonal construct, trust has been applied to the relationship between automated systems and human operators. Trust can be defined as, "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer, Davis, & Schoorman, 1995, pg. 712). Placing an appropriate amount of trust in an automated system is critical (Parasuraman & Riley, 1997). Misuse and disuse of automation, defined as over and under reliance respectively, can be attributed to over or under trust in the automated system at least partially (Parasuraman & Riley, 1997). Appropriate calibration of trust in accordance with system performance can aid in preventing misuse and disuse (Merritt, Lee,

Unnerstall, & Huber, 2014). The visual displays of reliability used in this study, aim to assist drivers in calibrating their trust appropriately.

Trust is often measured subjectively through multi-item scales. The Trust in Automation Scale, a 12-item scale, was developed to measure trust in an automated system (Jian, Bisantz, & Drury, 2000). Through the development of the scale, the authors found that there was no need to have separate measures for distrust and trust due to their high negative correlations, thereby indicating that trust and distrust are anchors for the spectrum of trust rather than two different underlying constructs. However, the authors have indicated that there could be insights from keeping the subscales of trust and distrust separate for analysis (Jian et al., 2000).

1.5 Dynamic displays

Static displays are those in which the information being displayed over time does not change (Wogalter & Laughery, 2012). Dynamic displays are those in which the information displayed to a user changes over time (Sanders & McCormick, 1993). These types of displays are very common in complex systems such as vehicles.

There are three categories of dynamic displays: quantitative, qualitative, and representational. Quantitative displays are those in which the variable of interest's exact value is displayed to the user either through a digital display or through analog means. These types of displays are best for when exact judgments must be made based on the variable of interest or the precise value is important for completing the task safely. Qualitative displays show variables in ways that are not directly numeric. For example, the battery display on a cell phone is a qualitative display as is shows the current battery level

by adjusting the fill of the battery over time. These displays are best for making relative judgments over time or between displays such as ensuring that the oil pressure in a car is within an acceptable range. The last category, representational displays, display the information to the user in a more abstract fashion. In visual displays, this could be in a graphical form (Sanders & McCormick, 1993). Representational displays are often used to increase situation awareness in the context. It is important to ensure that these displays match the operator's mental model or understanding of a system in order for them to be effective (Bennett, Nagy, & Flach, 2012; Endsley, 1998; Sanders & McCormick, 1993).

Representing the system that the display depicts in a way that allows for coherence and correspondence is essential to display design. Coherence is described as ensuring that the display matches the mental model of the user to be effective. If the display does not match the user's expectations of what it should look like or their understanding of the system that it represents, it could only serve to confuse the user and lead to errors. Correspondence, on the other hand, is the information from the domain or system that needs to be displayed for the operator to complete his or her tasks. Without the necessary information required, the operator would be unable to perform their tasks. The interaction between correspondence and coherence is the display, the operator's window to the system (Bennett et al., 2012).

1.5.1 Automation uncertainty displays

Dynamic displays of automation uncertainty, or the reliability of the automated system, have been shown to improve trust calibration and the ability to take control of the vehicle when necessary (Heldin, Falkman, Riveiro, & Davidsson, 2013). These displays

have also been shown to increase time to collision in highly automated vehicles (Beller, Heesen, & Vollrath, 2013). Even at lower levels of automation (e.g. adaptive cruise control), automation reliability displays have been shown to enhance reactions to failures in automation (Seppelt & Lee, 2007). However, these studies have not focused on the information content of the displays such as whether they present quantitative, qualitative, or representational information. This research will add to the existing literature through examining the role of information type in the effectiveness of reliability displays.

1.6 Current study

There are two primary objectives for this research. First, this study will determine how handover scenarios impact workload, driving performance, trust, and SA. Second, three types of automation reliability displays will be tested to determine if they increase SA and driving performance in handover scenarios while maintaining or reducing workload and aiding in appropriate trust calibration.

Specifically, this study plans to answer these research questions:

RQ1: What is the impact of handover on workload?

RQ1a: Do automation reliability displays reduce workload due to handover?

RQ2: What is the impact of handover on SA?

RQ2a: Does the addition of automation reliability displays improve SA?

RQ2b: Does this change in SA improve driving performance in handover and improve safety of automated to manual driving transitions?

RQ3: How does automation failure impact trust in automation?

RQ3a: How does the addition of automation reliability displays impact trust in automation?

CHAPTER 2. METHODS

The experimental design of this study consisted of two independent variables: automation failure (within subjects) and displays (between subjects). The displays manipulation had four levels: (1) no display; and reliability displays using (2) quantitative information (percentage of reliability); (3) qualitative information (direct representation of a number); and (4) representational information (abstract representation of a number). There were five dependent variables: gaze distribution; objective workload (measured by pupil size and heart rate); subjective workload (measured by NASA-TLX); trust (measured by the Trust in Automation Scale (Jian et al., 2000)), and driving performance (measured by following distance and lane deviation). SA was measured both subjectively and objectively. The SART was used to gain insight into the operators perceived SA during the task. Complimentarily, an implicit performance measure, following distance, gives added objectivity to SA measurement.

2.1 Participants

Participants in this study were Georgia Institute of Technology psychology students who were enrolled in the SONA system. They received course credit for their participation. Inclusion criteria for participating in this study included: normal to corrected normal vision, mobility, and hearing, a valid driver's license, and a minimum two years of driving experience. The experience requirement ensured that all participants had adequate driving experience to prevent novice drivers from impacting results.

Eighty-six participants were recruited to participate in the study. However, only 62 participants' data were analyzed for the purposes of this experiment. Some participants were excluded due to taking over the automation during the baseline drive, others were excluded for technical or administrative failures, and lastly, the majority of participants were excluded due to taking over control of the automated vehicle prior to failure in the handover drive.

The final group of participants (N=62) ranged in age from 18 to 24 years old (M = 19.6 years, SD = 1.26) and had an average of 3.24 years of driving experience (SD=1.11). There were 25 females (40.3%) and 37 males (59.7%) who participated in the study. Participants were moderately familiar with automated safety systems prior to participating in this study. As seen in Table 1, participants primarily identified with being familiar with safety features (35.5%) and some participants had direct experience driving with automated safety features (30.7%) through either owning a vehicle with automated safety features or driving one previously.

	Frequency	**Percent**
"I own a vehicle with one or more automated safety features"	6	9.7
"I have driven a vehicle with one or more automated safety features."	13	21
"I have been a passenger in a vehicle with one or more automated safety features."	12	19.4
"I am familiar with automated safety features"	22	35.5
"I have never heard of automated safety features prior to participating in this study"	9	14.5

Table 1. Participant self-reported prior experience with automated safety features.

2.2 Materials

2.2.1 Driving environment

2.2.1.1 Driving task

Participants completed two drives in the MiniSim Version 2.2.1. Both took place on a rural, two-lane, curvy road with low to moderate traffic, seen in Figure 1 and Figure. Participants were instructed to maintain the initial following distance with the vehicle in front of them (50 feet). They drove with automated lane keeping, a system that maintains vehicle position within the lane, turned on initially for both drives. The route for the baseline drive is seen in Figure 1. Participants started at point A and drove until they reached point B, which took approximately 6 minutes. For the handover drive (Figure 2), participants began at point A and ended at point E approximately. The handover scenario ended based on elapsed time rather than location which is why the end point is approximate. If the participant was in a display condition, the reliability level reduced from high (starting) to moderate at point B and from moderate to low at point C. At point D, all participants experienced the automation failure requiring manual takeover of the vehicle's steering.

Figure 1. Route for baseline drive.

Figure 2. Route for the handover drive.

2.2.2 Trust in Automation

Participants completed the Trust in Automation scale (APPENDIX G. Trust in Automation Scale) twice throughout the experiment, once after each driving segment (Jian et al., 2000).

2.2.3 Performance measures

Following distance was used as a measure of SA. This performance measure gave insight into the level of situation awareness the driver has as it is their primary task to maintain 50 feet of distance from the vehicle they are following. Variance of following distance was used as a dynamic measure of to determine SA. The follow vehicle had a dynamic speed that had an average of 50mph but could range between 40mph and 60mph forcing participants to adjust their speed throughout the drive to maintain the same distance. The speed of the follow car was random and normally distributed over the course of the drive as to maintain the set minimum, maximum, and mean.

Lane deviation served as a driving performance measure. This is the difference from the vehicle's lateral position to the center of the lane. When adaptive cruise control is on, this value will be zero. However, once the automation failure occurs and participants must regain control of the vehicle, the lane deviation will be a measure of how smoothly the participant transitioned from automated to manual driving modes. If there is low lane deviation during handover, this will indicate smooth transition from automated to manual; higher levels of lane deviation will indicate more difficulty in the transition to manual driving.

2.2.4 Situation awareness measures

Participants completed SART twice during the experiment to assess their subjective SA (Taylor, 1990). The first rating was collected after completion of the first drive and the second, after the second drive. In addition, driving performance measures served as implicit performance measures of SA. Two subtasks specifically gave insight into the driver's awareness of the driving environment. Following distance, from the vehicle they are instructed to follow at a given distance prior to beginning the drive, served as a dynamic measure of SA throughout the driving task. Increased variance in following distance indicates decreased SA; whereas, decreased variance in following distance will indicate enhanced SA. As this was the primary driving task, the degree to which participants can adhere to specified guidelines indicated their knowledge of their own vehicle's location and speed relative to the follow vehicle's location and speed.

In addition to subjective and performance measures of SA, eye tracking data was collected. Gaze distribution was calculated for both drives of the experiment based on the

variance of gaze location throughout each drive. Spread across the driving simulator and displays, indicates broad attention allocation and thereby information acquisition throughout the driving environment. Narrow gaze distribution indicates perceptual tunnelling and lack of information acquisition throughout the environment.

2.2.5 Workload measures

2.2.5.1 Subjective measures

NASA-TLX was used to measure subjective workload. This measure will be calculated with weights as described by the authors (Hart & Staveland, 1988). This was administered to participants twice, once after the completion of each drive.

2.2.5.2 Physiological measures

A fixed, three-camera SmartEye eye-tracking system will be used in this study to collect both pupil size and gaze location throughout both drives based on calibrated settings for each participant.

Heart rate data was collected via a three-lead system as shown in APPENDIX I. HEART RATE MONITOR PLACEMENT INSTRUCTIONS. This data was analysed to assess changes in heart rate due to workload over the course of the experiment.

2.2.6 Reliability displays

In a preliminary study, twelve reliability displays in three different categories: quantitative, qualitative, and representational (four displays per category) were developed by designers (Noah, Gable, Chen, Singh, & Walker, 2017). A card sorting study was

completed to determine whether mental models of the designers matched those of participants. The participants were 50 Georgia Tech undergraduate psychology students. The results showed that the displays in the figures below (Figure 3, Figure 4, and Figure 5) were matched with highest accuracy by participants across the three display types (Noah et al., 2017).

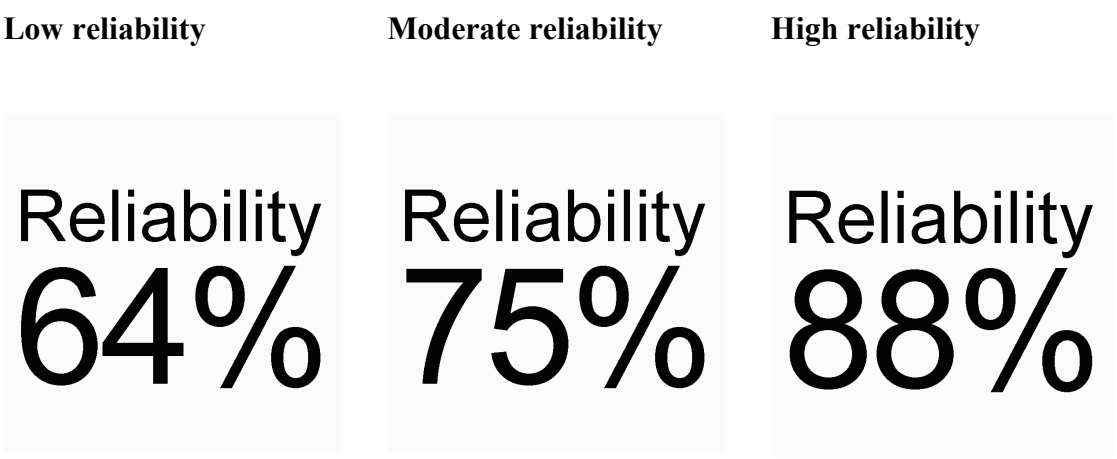


Figure 3. Quantitative displays of reliability.

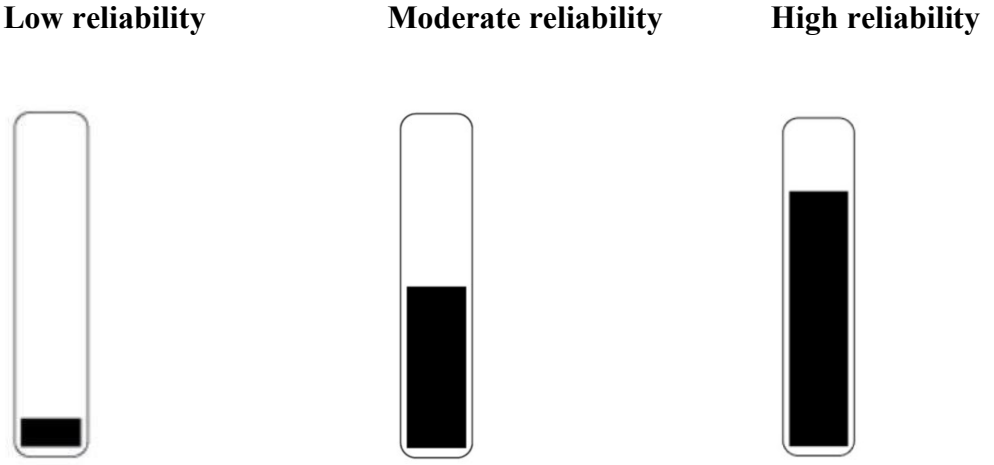


Figure 4. Qualitative displays of reliability.

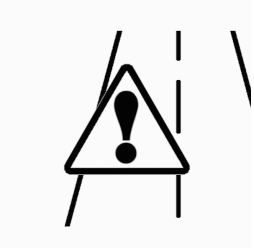
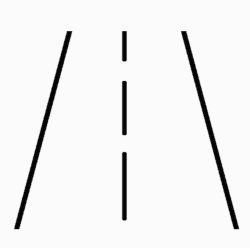
Low reliability	Moderate reliability	High reliability
		
*Roadway moves from right to left behind the triangle.	*Roadway markings move up and down slightly behind the triangle.	*Roadway markings move up and down behind the triangle.

Figure 5. Representational displays of reliability.

2.3 Procedure

Participants were randomly assigned to a condition: no display (n=14), quantitative (n=16), qualitative (n=17), or representational (n=15). Upon completing the informed consent form and filling out a demographic questionnaire (APPENDIX C. Demographics questionnaire) participants completed Georgia Tech Simulator Sickness Screening protocol to ensure that they would not suffer from any physical discomfort while completing the experiment (Gable & Walker, 2013). Next, participants completed the baseline drive. This initial drive served as a training route for participants. They became familiar with the display if they are in a display condition, as well as familiar with the automated lane keeping system. This drive also served as a baseline for dependent variables. Upon completing the first drive, participants completed NASA-TLX, SART, and the Trust in Automation scale. Participants began the second drive (handover) with

automated lane keeping turned on. Throughout the drive, participants in the display conditions were shown decreasing reliability of the system and experienced a failure of the automated lane keeping system. Those participants in the control condition, where no display was present, did not receive the reliability information but experienced the same failure. Upon experiencing this automation failure, participants will be required to take over control of the vehicle to continue to drive for approximately 7-8 minutes. Upon completing the driving task, the participants completed the Trust in Automation scale, NASA-TLX, and SART once more. They also completed a brief questionnaire about their experience during handover (APPENDIX I. Handover Experience Questionnaire). Finally, participants were debriefed (APPENDIX J. Debrief), released, and given credit for their participation.

CHAPTER 3. RESULTS

3.1 Results overview

After a description of the handover task in terms of the eye tracking and heart rate measures collected, this section will follow the same ordering as the introduction and will be presented in terms of the dependent variables measured. Several analyses were completed to test each of the hypotheses. Bonferroni corrections to alpha levels were used when appropriate to adjust for family-wise Type 1 error.

3.2 Handover task characterization

To better characterize the dynamic task of automation handover, heart rate, pupil size and gaze distribution data were collected throughout the baseline and handover drives. This data allows for characterization of workload and attention allocation throughout the drives as the task dynamics change due to automation failure. The means reported in this section are for all participants, not just those who have baseline and handover drive data. Participants without baseline and handover data were not used for the inferential analyses so the means reported in this section will be different than those used in the inferential analyses. As this section is aimed at providing a description of how the heart rate, pupil diameter, gaze distribution change throughout the drives, all possible participants were included.

3.2.1 Heart rate

Figure 6 shows average heart rate across all participants throughout the baseline and handover drives. The automation handover during occurred at approximately 7 mins and 15 seconds into the drive (435 seconds), marked by a vertical line in Figure 6.

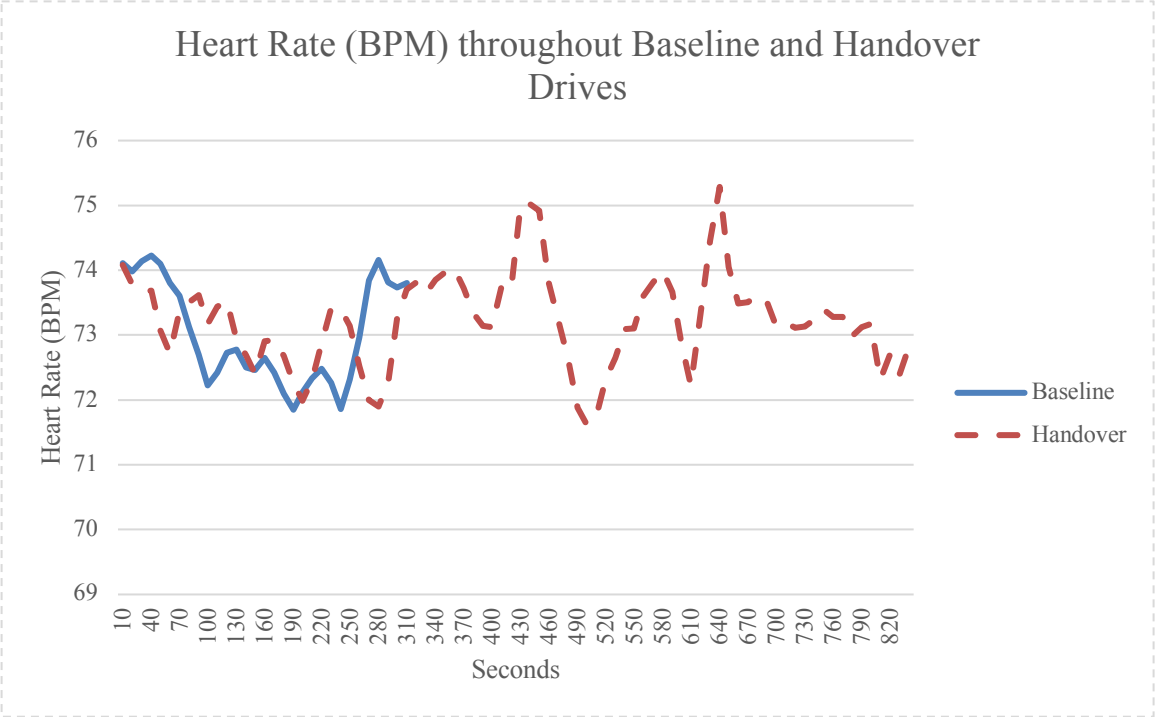


Figure 6. Average heart rate throughout the baseline and handover drives.

To further understand how workload, as measured by heart rate, changes over time the handover drive was broken into segments as seen in Table 2 and Figure 7. The first phase of the handover drive, with the automated lane keeping turned on, is the automated handover drive or segment A. The second phase of the handover drive is the control transition. This was at approximately 435 seconds but changed by participant as the automation failed at the same location in each driving scenario, rather than at the same time, to ensure similar experiences between participants. The final phase of the drive is the

manual driving portion and is marked as segment C in Figure 7. The average heart rate for each of these segments can be seen in Table 2.

Drive Segment	**Average HR (BPM)**
Baseline Drive	72.94
Handover Drive (entire drive)	73.34
Automated Handover Drive (A)	73.17
Control transition (B)	74.99
Manual Handover Drive (C)	73.51

Table 2. Average heart rate (HR) in beats per minute (BPM) for each driving segment.

For segment A of the handover drive, heart rate remained steady with a slight increase at the end of the segment leading into the control transition. For the manual phase of the drive (segment C), the heart rate generally trended down after settling into manual driving about halfway through the phase. There is a slight increase towards the end of the drive likely due to the sharp turn seen near the end of the route in Figure 2. Therefore, the participants seemed to experience a steady amount of workload in the automated driving phase, followed by a phase of higher workload for the control transition, and finally a downward trending workload about halfway through the manual driving phase.

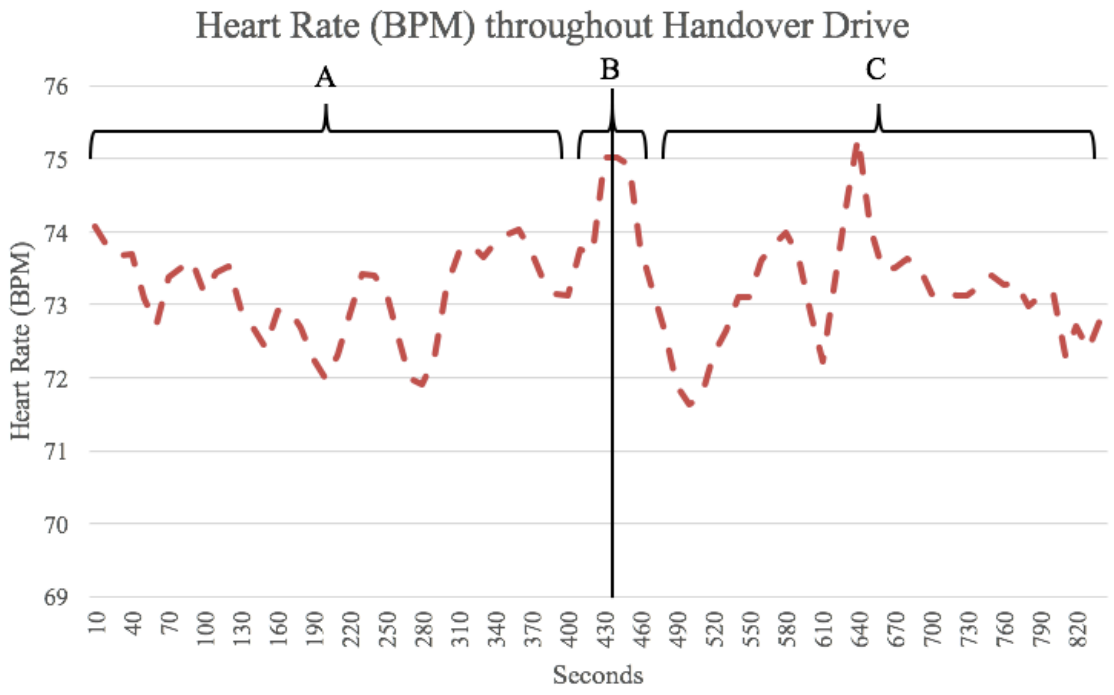


Figure 7. Segmented average heart rate for handover drive.

3.2.2 Pupil diameter

Figure 8 shows the average pupil diameter throughout the baseline and handover drives. For the baseline drive, the pupil diameter starts high and trends downward as participants gained familiarity with the automated lane keeping system and the simulator. The upward trend at the very end of the baseline drive is likely due to the curve at the end of the drive as seen in Figure 1. Therefore, there's a high workload experienced at the beginning of the baseline drive that decreases with system familiarity. To further understand the handover drive, the same segmentation procedure was used as with the heart rate data. This allows for comparison of the automated phase of the drive, the control transition phase, and the manual driving phase (Segments A, B, and C respectively, Figure 9). Average pupil diameter for each of these phases can be seen in Table 3.

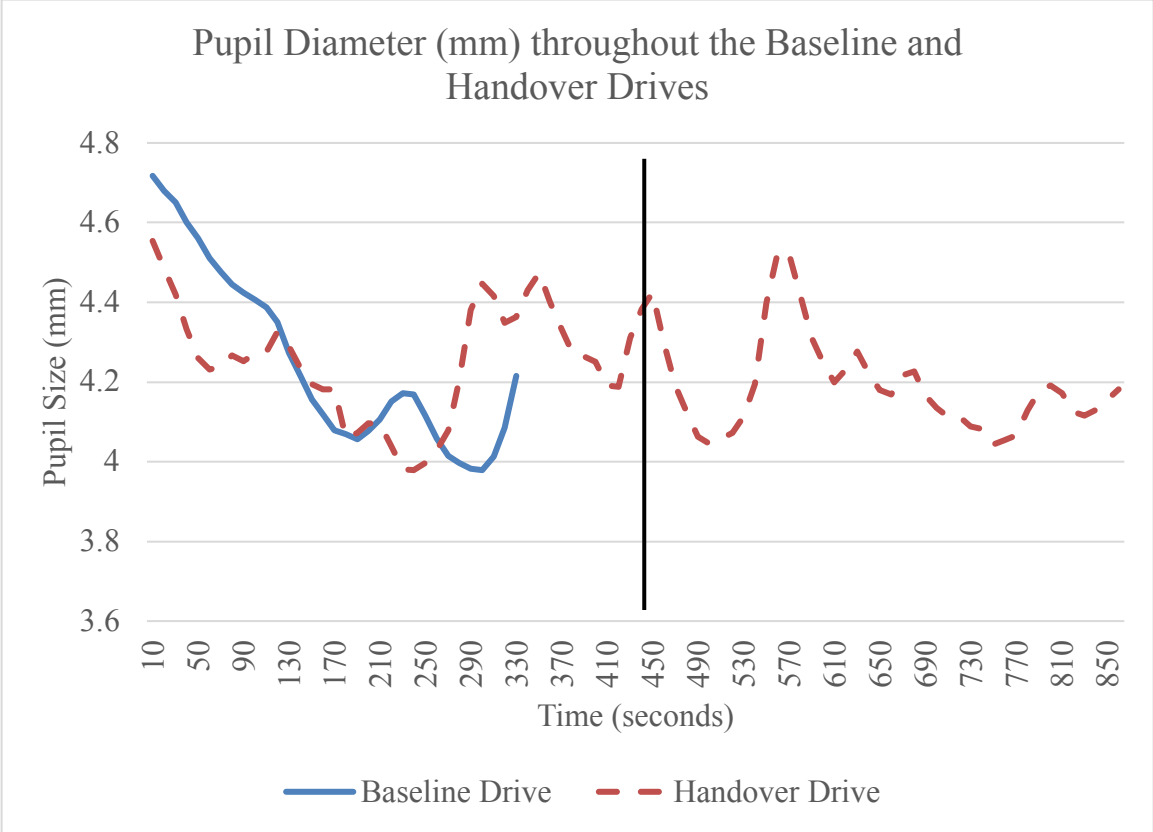


Figure 8. Pupil size throughout the baseline and handover drives.

Drive Segment	**Average Pupil Diameter (mm)**
Baseline Drive	4.25
Handover Drive (entire drive)	4.31
Automated Handover Drive	4.32
Control transition	4.41
Manual Handover Drive	4.29

Table 3. Average pupil diameter (mm) in each driving segment.

In the handover drive, the participants experienced a similar pattern of workload in the automated phase (segment A, Figure 9) as in the baseline drive with a high starting

workload and general downward trend until the second half of the phase. The latter half of the automated driving phase is characterized by more roadway curvature and changing levels of reliability (for the display conditions) which leads to higher workload (larger pupil diameter). The control transition phase, segment B, has a peak in pupil diameter and therefore, workload. For the manual driving phase, there is an initial increase in workload followed by a decreasing trend. The very last portion of the manual driving phase has a sharp turn which is seen as higher workload (larger pupil diameter) towards the end of the phase.

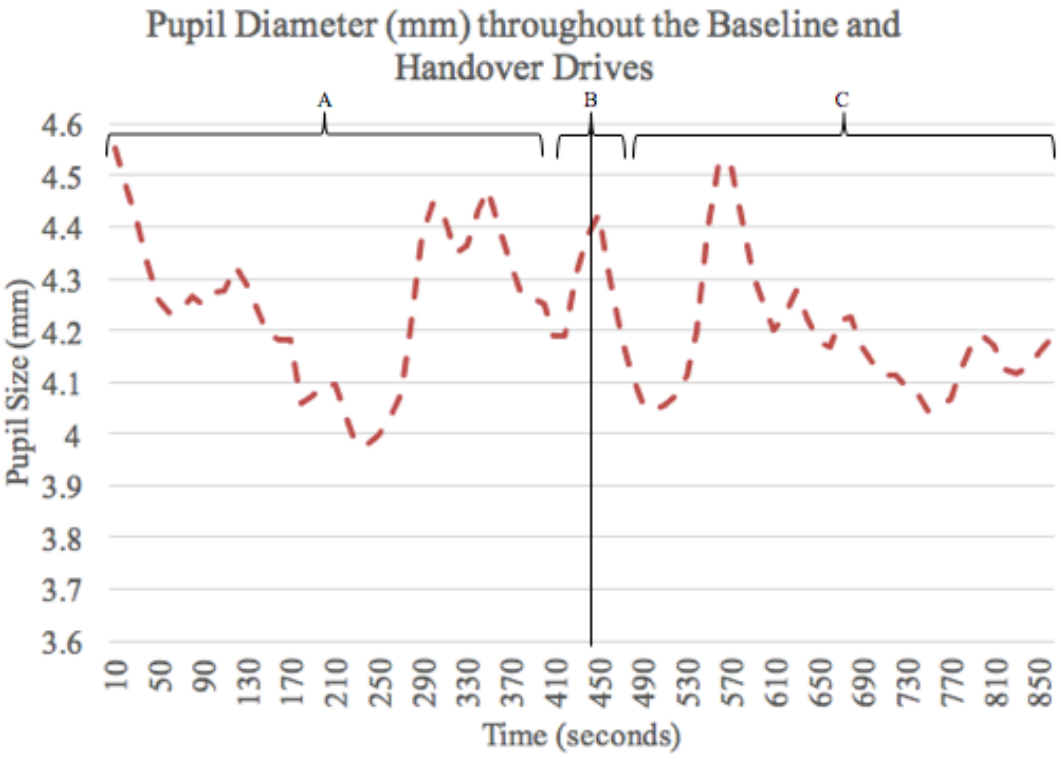


Figure 9. Segmented average pupil diameter (mm) for the handover drive.

3.2.3 Gaze distribution

Gaze distribution was studied in both the horizontal and vertical directions to understand how gaze distribution changed bi-directionally throughout the baseline and handover drives.

3.2.3.1 Horizontal gaze distribution

The average variance of horizontal gaze position throughout the baseline and handover drives can be seen in Figure 10. For the baseline drive, there is a slight increase in average variance in horizontal position throughout. This slight increase indicates less perceptual tunnelling due to reduced workload as familiarity increases. The vertical line separating the handover drive indicates the average time point at which participants experienced the automation failure. To further characterize the horizontal gaze position in the handover drive, it has been further divided into three phases (A, B, and C) as seen in Figure 11. Average variance of each segment can be seen in Table 4.

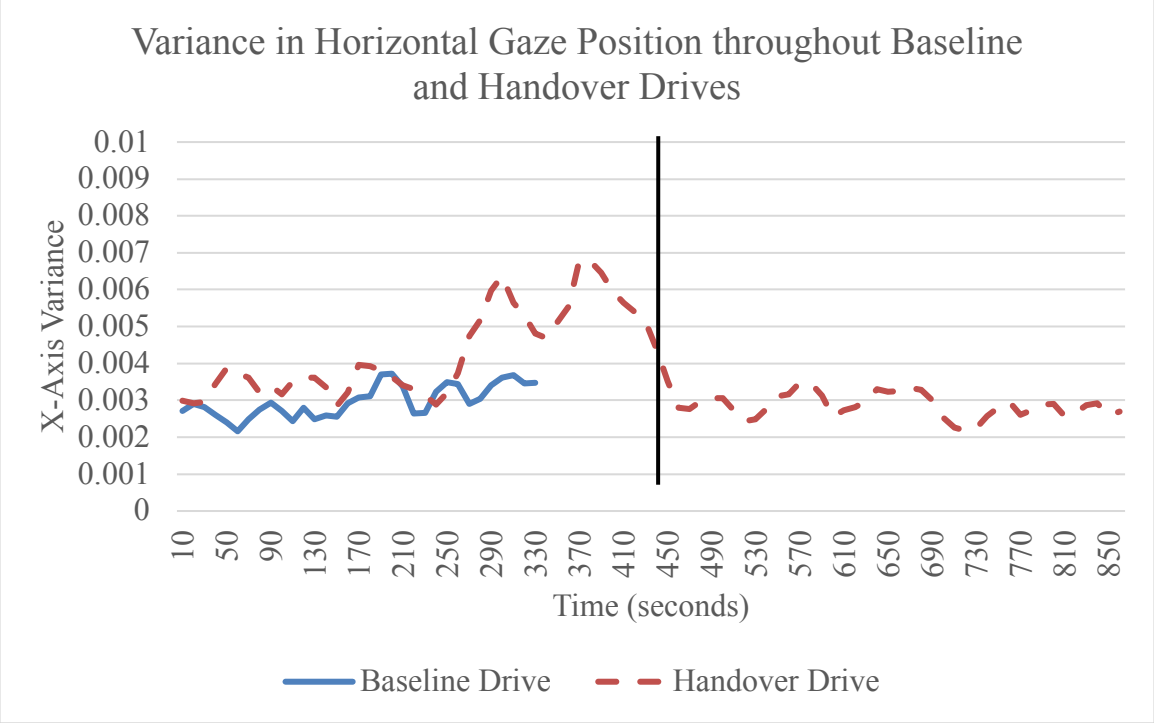


Figure 10. Variance in horizontal gaze position throughout the baseline and handover drives.

Drive Segment	**Average Horizontal Position Variance**
Baseline Drive	0.00298
Handover Drive (entire drive)	0.00359
Automated Handover Drive	0.00422
Control transition	0.00498
Manual Handover Drive	0.00287

Table 4. Average horizontal variance for each driving segment.

In segment A, the automated portion of the handover drive, there is a similar trend to what was seen in the baseline drive. Segment A, like the baseline drive, is characterized by increasing variance in horizontal gaze distribution due to decreasing workload as task familiarization increases. Segment B, the control transition phase, is characterized by a reduction in average horizontal gaze position variance due to higher workload from the manual driving task. This reduced variance persists throughout segment C indicating perceptual tunneling.

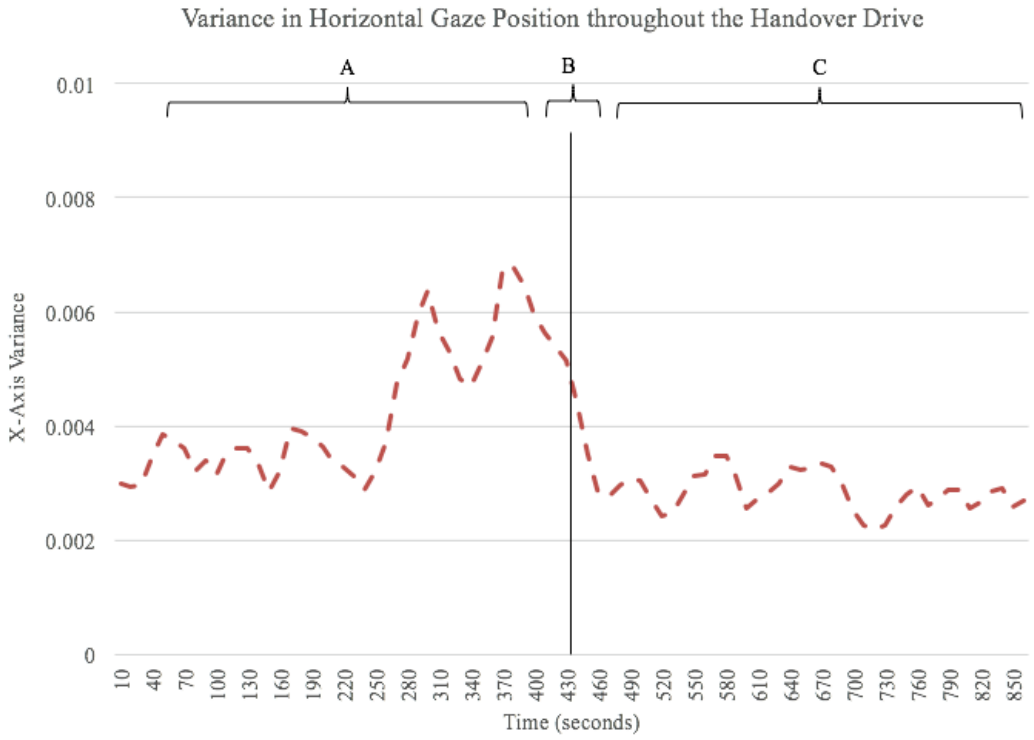


Figure 11. Segmented average variance in horizontal gaze position throughout the handover drive.

3.2.3.2 Vertical gaze distribution

Trends in the baseline and handover drives for vertical gaze distribution are very similar to horizontal gaze distribution trends. The baseline drive is characterized by slightly increasing average variance in vertical gaze distribution. This indicates that participants were increasingly scanning the environment with higher levels of familiarity, showing no perceptual tunneling effect. The handover drive is generally characterized as automated and manual driving segments which are separated by the vertical line in Figure 12. Further characterization of this drive is possible with further division into three segments as seen in Figure 13.

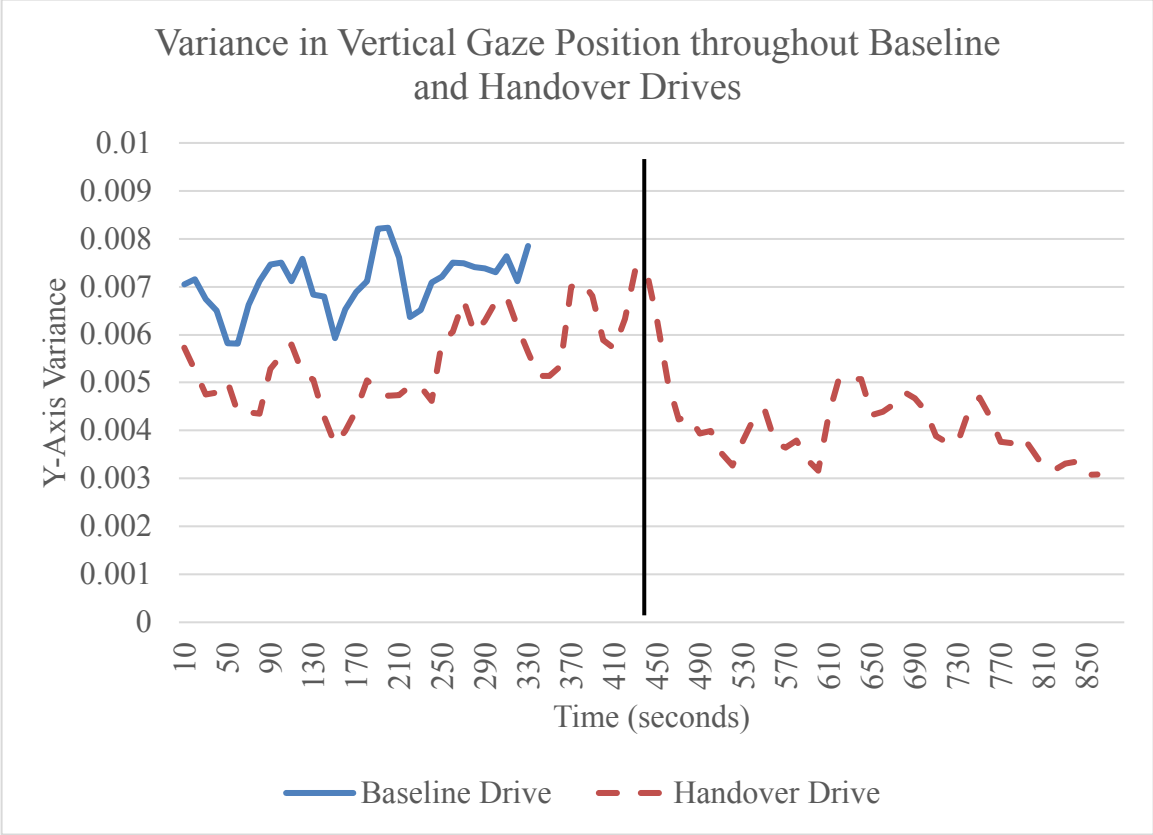


Figure 12. Variance in vertical gaze position throughout the baseline and handover drives.

Drive Segment	**Average Vertical Position Variance**
Baseline Drive	0.00707
Handover Drive (entire drive)	0.00481
Automated Handover Drive (A)	0.00531
Control transition (B)	0.00705
Manual Handover Drive (C)	0.00416

Table 5. Average vertical variance for each driving segment.

The average variance in vertical gaze position for the handover drive begins much the same as the baseline. In segment A, there is a general increase in vertical gaze position

variance as workload decreases and familiarity increases. This increase continues through the control transition, segment B, and then variance rapidly decreases in the manual driving phase, segment C. This rapid decrease in variance during the manual driving phase is due to higher workload due to the nature of the non-automated driving task and is evidence for perceptual tunneling.

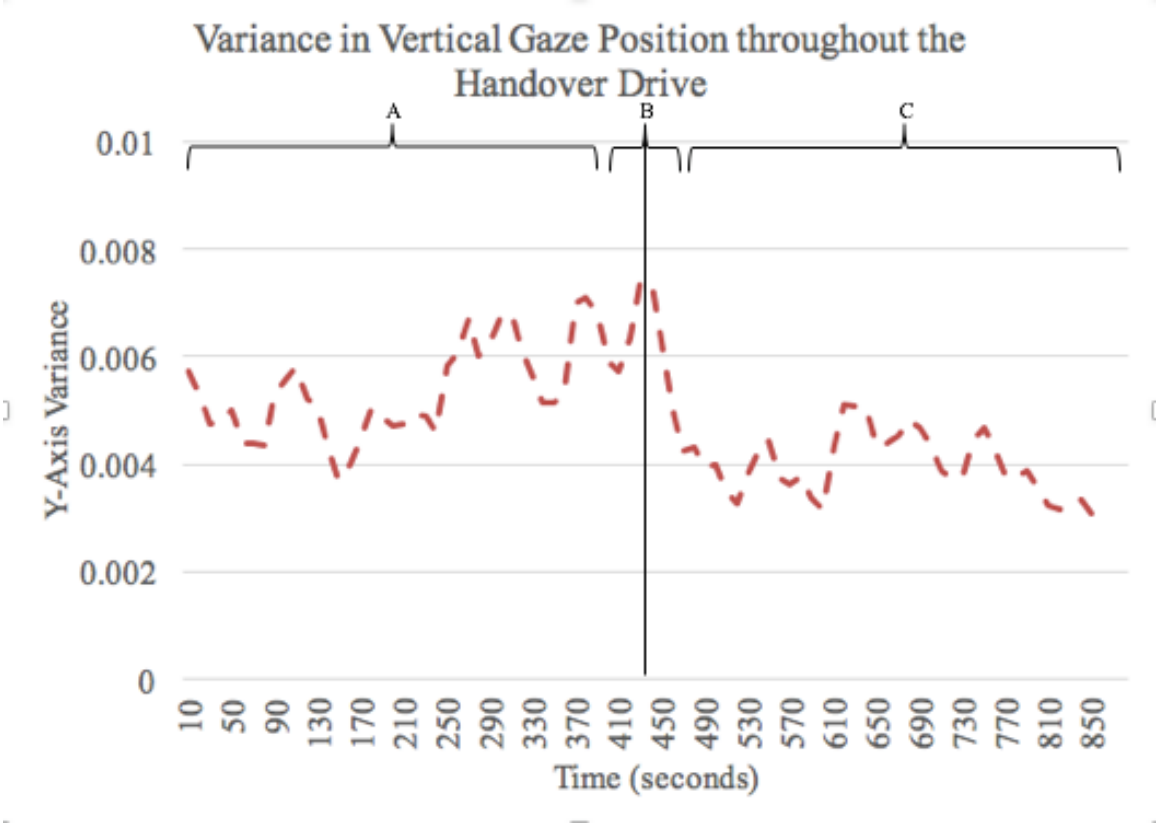


Figure 13. Segmented average variance in vertical gaze position throughout the handover drive.

3.2.4 Task characterization summary

The baseline drive can be generally characterized as a familiarization task. At the onset of the task, participants experience an initial high workload that generally reduces throughout the drive with a slight increase at the end due to road curvature. This workload

reduction allows for a wider horizontal and vertical gaze distribution as more resources are available for visual attention throughout the task.

The automated portion of the handover drive is very similar to the baseline. There is an initial high level of workload that then reduces through the first half of the phase. In the latter half of the automated driving phase, there is an increase in workload due to a combination of increased road curvature and a reduction reliability level for those in the display conditions. The transition of control from automated to manual lane keeping is characterized by an increase in workload and a decrease in gaze distribution. The first half of the manual driving segment remains at a higher workload with a general reduction in the second half. The sharp turn at the end of the handover drive causes an increase in workload at the very end of the drive. Gaze distribution for the manual driving phase is decreasingly variable in comparison to the automated and control transition phases. This decrease in variance, a decrease in overall scanning, is evidence for perceptual tunnelling.

3.3 Cognitive workload

3.3.1 Hypothesis 1a

1a: Cognitive workload will be lower in the baseline drive, with no failures, than the handover drive in which failures occur.

3.3.1.1 NASA-TLX

To test this hypothesis, a 2 (Drive) x 4 (Condition) split-plot ANOVA was used. There was a significant difference between baseline and handover workload as measured by the weighted total of NASA-TLX, $F(1, 58) = 84.570$, $p < .001$. This result showed that

across condition, workload was perceived to be lower in the baseline drive than in the handover drive. There was no significant difference in workload by condition, $F(3, 58) = 0.129$, $p=.943$. This suggests that higher subjective workload was experienced in the handover drive ($M = 51.82$, $SD = 16.62$) than in the baseline drive ($M = 36.973$, $SD = 12.58$).

3.3.1.2 Heart Rate

Prior to the analyses for this hypothesis, the handover drive heart rate data was analyzed to determine if there was a significant difference in average heart rate between segments. The results of this test informed the analyses chosen to test the hypothesis. A 3 (Drive Segment) x 4 (Condition) split-plot ANOVA was used to test the effect of drive segment and condition on heart rate. A Hyunh-Feldt correction was used for the within subjects effects to correct for sphericity. The results showed that there was a significant main effect of drive segment, $F(1.938, 81.416) = 3.299$, $p=.043$. There was no significant interaction effect, $F(5.815, 81.416) = 0.402$, $p=.871$. These results suggest that there is a difference in average heart rate based on the driving segment.

A set of three 2 (Drive) x 4 (Condition) split-plot ANOVAs were completed to determine if there were differences between the baseline and any of the handover driving segments. The first ANOVA tested the baseline drive and the automated segment of the handover drive. This analysis resulted in no significant differences due to drive, $F(1, 36) = 0.428$, $p=.516$, or condition, $F(3,36) = 1.696$, $p=.185$. There was also no significant interaction effect $F(3,36) = 0.238$, $p=.869$. Next, a test of the baseline drive compared to the control transition phase of the handover drive was completed. The results for this

ANOVA were also non-significant for drive, $F(1, 36) = .994$, $p=.325$, and condition, $F(3,36) = 1.884$, $p=.150$. There was also no interaction effect $F(3,36) = 0.178$, $p=.910$. Lastly a comparison of the baseline to the manual phase of the handover drive was completed. This analysis also yielded null results. There were no significant differences due to drive, $F(1,36) = .029$, $p=.865$, or condition $F(1, 36) = 1.792$, $p=.166$. There was also not a significant interaction effect, $F(3,36) = 0.350$, $p=.789$. These results show that there was no significant effect of to drive or condition on average heart rate.

3.3.1.3 Pupil Diameter

Prior to comparing baseline and handover drives, a comparison of the handover drive segments was completed with a 3 (Drive segment) x 4 (Condition) split-plot ANOVA to determine if the overall average of the handover drive should be used for analysis or if the segments should be used. The Huynh-Feldt adjustment for sphericity was used in this analysis. The results of this test showed no significant differences based on drive segment $F(1.082,41.099) = 0.724$, $p=.410$. Therefore, there was no significant difference in pupil diameter between driving segments. Further, no significant interaction between drive segment and condition was found, $F(6,41.099) = .813$, p=.502. The null results of this analysis led to testing the hypothesis with a simple comparison of average baseline pupil size and average handover pupil size.

ANOVA A 2 (Drive) x 4 (Condition) mixed ANOVA was used to understand mean differences due to drive and display condition with the pupil diameter data. The results of this test showed that there were no significant differences between the baseline and handover drives, $F(1, 21) = 3.665$, $p=.069$. No significant differences were found due to

display condition, $F(3, 21) = 0.065$, $p=.978$. There was also no significant drive by condition interaction effect $F(3, 21) = 0.686$, $p=.570$. Therefore, there was no difference in workload, as measured by pupil size due to drive or display condition.

3.3.1.4 Summary

The results of the statistical analyses suggest that there were no significant differences between the baseline and handover drives objectively (as measured by heart rate and pupil size data). However, there were significant differences found with the subjective measure of workload, NASA-TLX, that suggest higher workload was experienced in the handover drive than in the baseline drive. Therefore, hypothesis 1a is confirmed only for subjective workload.

3.3.2 *Hypothesis 1b*

During the handover drive, cognitive workload will be higher in the control (no display) condition than the display conditions.

3.3.2.1 NASA-TLX

In a 2 (Drive) x 2 (Display presence) mixed ANOVA, there was a significant main effect of drive, $F(1, 60) = 56.632$, $p<.001$. There was no significant main effect of display presence, $F(1, 60) = 0.117$, $p=.734$, and no interaction effect $F(1, 60) = 0.157$, $p=.694$. Therefore, there was no significant difference of workload found between the display and no display conditions. The displays did not have any mediating effect on workload.

3.3.2.2 Heart rate

To determine if there was a significant difference between the control, no display, and display conditions, a set of three 2(Drive) x 2(Display presence) ANOVAs were completed to account for differences in heart rate due to driving segment. First, a comparison of the baseline and automated phase of the test drive was completed. There was no significant difference found between the baseline and test drives $F(1,38) = 0.355$, $p=.555$. There was no significant difference due to display $F(1,38) = 1.339$, $p=.254$. There was also no interaction effect $F(1,38) = 0.002$, $p=.967$. Next, a comparison of baseline and the handover segment of the test drive was completed. There was no significant difference between drives, $F(1,38) = 0.834$, $p=.367$, or condition $F(1, 38) = 1.410$, $p=.242$. There was also no interaction effect $F(1, 38) = .026$, $p=.872$. Finally, a comparison of the baseline and manual drives was conducted. This analysis yielded no significant effects due to drive, $F(1,38) = 0.126$, $p=.725$, or condition, $F(1,38) = .980$, $p=.328$. There was also no interaction effect, $F(1,38) = 0.334$, $p=.567$.

3.3.2.3 Pupil Diameter

A comparison of the control and display groups was completed through a 2 (Drive) x 2 (Display presence) split-plot ANOVA. The results of this analysis showed that there were no differences between the drives, $F(1,23) = 1.269$, $p=.272$, display presence, $F(1,23) = .166$, $p=.687$, or an interaction $F(1,23) = 2.065$, $p=.164$. Therefore, there was no difference due to simply having a display of any kind versus no display in workload as measured by pupil diameter.

3.3.3 Hypothesis 1c

Representational displays will have reduced workload during the handover drive compared to quantitative and qualitative displays.

Hypothesis 1c was not evaluated as there were no significant differences in workload due to condition in the omnibus ANOVAs tested in Hypotheses 1a and 1b.

3.3.4 Hypothesis 1d

Participants in the display condition will return to their baseline workload quicker than those in the control condition after the handover occurs.

This hypothesis was not evaluated. The test for heart rate differences between driving segments was significant; however, individually, there were very few participants within each group that experienced an increase in heart rate during the handover segment and also had baseline heart rate data for comparison. For the pupil diameter data, this was not evaluated as there were no significant differences across driving segments.

3.3.5 Workload summary

These results suggest that there was a difference in subjective experience between the baseline and handover drives, as measured by NASA-TLX; however, the objective measures of workload did not capture these changes. The objective measures suffered from technical failures that led to smaller sample sizes for the analyses which impacted these results. Additionally, there were no differences in workload due to condition seen in either the objective or subjective measures.

3.4 Situation Awareness

3.4.1 Hypothesis 2a

SA (as measured by SART and gaze distribution) will be highly positively correlated.

A correlation of the baseline gaze distribution measures and SART scores revealed that there is not a significant correlation between baseline SART scores and gaze distribution, horizontal: r=0.122, p=.473; vertical: r=.147, p=.384. For the handover drive, there was first an overall comparison made between the average horizontal and vertical gaze position throughout the entire drive and then based on driving segments. Neither horizontal or vertical gaze distribution was found to be a significant correlate of the handover drive SART scores, horizontal: r=-0.041, p=.789; vertical: r=0.036, p=.822. For the automated portion of the test drive there was no significant correlation with either the horizontal, r=-0.104, p=.510, or vertical, r=-0.139, p=.380 gaze distribution and SART. Likewise, there was no significant correlation between the handover SART scores and gaze distribution in the handover portion of the drive, horizontal: r = -0.103, p=.507; vertical: r=-0.058, p=.716. Finally, there was no significant correlation between the manual driving phase of the handover drive and the baseline drive, horizontal: r=0.094, p=.553; vertical: r= 0.089, p=.575. These findings suggest that there is no relationship between the SART scores and the gaze distribution measures in either the baseline or handover drive.

3.4.2 Hypothesis 2b

SA (as measured by SART and gaze distribution) will be higher, on average, in conditions with a display than in those without a display.

3.4.2.1 SART

A 2 (Drive) x 4 (Display Type) split-plot ANOVA from the SART scores, a significant difference was found between baseline and handover drives, $F(1, 58) = 15.228$, $p<.001$. There was no significant difference between conditions, $F(3, 58) = 0.742$, $p=.531$. Likewise, there was no significant interaction effect, $F(3, 58) = 1.153$, $p=.335$. This result suggests a significantly higher perceived SA in the baseline drive ($M = 17.26$, $SD = 5.411$) than the handover drive ($M = 14.71$, $SD = 4.396$).

To follow this analysis, a 2 (Drive) x 2 (Display presence) split-plot ANOVA was completed. The results showed that there was a significant difference between the baseline and handover drives in terms of subjective situation awareness, $F(1, 60) = 16.461$, $p<.001$. There was no significant difference found due to the presence of the display, $F(1, 60) = 1.255$, $p=.267$, and no interaction effect $F(1, 60) = 2.492$, $p=.120$. Therefore, there was no significant difference in situation awareness due to the presence of a display.

3.4.2.2 Gaze distribution

Prior to testing the hypothesis, the gaze distribution data for the handover drive were analyzed to determine whether it differed significantly by drive segment. These results were used to determine how to test the hypothesis. A 3 (Drive Segment) x 4 (Condition) split-plot ANOVA was completed for both horizontal and vertical gaze distribution to determine if there are differences due to driving segment and if these differences result in an interaction with condition. A Huynh-Feldt correction for sphericity was used for the within subjects' effects.

For horizontal gaze distribution, there was a significant effect of driving segment $F(1.395, 53.013) = 8.756$, $p=.002$. There was no interaction effect $F(4.185, 53.013) =$

0.581, p=.686. There was also no main effect of condition $F(3, 38) = 0.800$, p=.502. This result suggests that the handover drive should not be analyzed as a whole for this measure. The hypothesis was tested with a set of ANOVAs in order to test each handover driving segment against the baseline drive.

For vertical gaze distribution, there was no significant difference due to drive segment $F(1.172, 44.525) = 1.716$, p=.198, or condition, $F(3,38) = 1.533$, p=.222. Finally, there was no interaction effect $F(3.515, 44.525) = 1.027$, p=.398. Therefore, there was no significant difference in vertical gaze distribution due to drive segment. For the vertical gaze distribution data, the hypothesis will be tested simply by comparing average baseline to average handover drive.

To test the hypothesis a 2 (Drive) x 4 (Condition) split-plot ANOVA was completed. A set of three ANOVAs was completed for horizontal gaze distribution and a single ANOVA was completed for vertical gaze distribution.

For horizontal gaze distribution, the first test compared the average variance in the baseline drive to the average variance of the automated driving section of the handover drive. This analysis yielded no significant differences due to drive, $F(1, 21) = 1.646$, p=.214, or condition, $F(3, 21) = 0.055$, p=.983. Likewise, there was no significant interaction effect, $F(3, 21) = 0.716$, p=.554. The second test compared the baseline and control transition phase of the handover drive. This test yielded no significant differences in horizontal gaze distribution due to drive, $F(1, 21) = 2.262$, p=0.148, or condition $F(3,21) = 0.280$, p=.839. There was also no significant interaction, $F(3, 21) = 0.849$, p=.183. Finally, a comparison of the manual portion of the handover drive and the baseline drive

was completed. This analysis yielded no significant differences due to drive, $F(1, 21) = 0.385$, $p=0.541$, or condition, $F(3,21) = 0.024$, $p=.995$. There was also no significant interaction effect, $F(3, 21) = 1.169$, $p=0.345$. These results show that there was no significant difference in horizontal gaze distribution due to the drive or display condition.

Further analyses of the horizontal gaze distribution data were conducted to determine if there were differences between the control (no display) and display conditions generally. To test this, a set of three 2 (Drive) x 2 (Display presence) ANOVA were conducted. First, a comparison of the baseline and automated phase of the test drive was completed. This analysis yielded no significant differences due to drive, $F(1, 23) = 3.917$, $p=.060$, or condition, $F(1, 23) = 0.177$, $p=.678$. There was also no significant interaction effect, $F(1, 23) = 2.013$, $p=.169$. Next a comparison of the baseline drive and control transition phase of the handover drive was completed. The results of this analysis showed that there is a significant difference between drives $F(1, 23) = 5.150$, $p=0.033$. No significant differences were found between conditions $F(1, 23) = 0.898$, $p=.353$. There was also no significant interaction effect, $F(1, 23) = 2.698$, $p=.114$. Finally, a comparison of baseline and manual driving phase of the handover was completed. This analysis revealed that there are no significant differences due to drive, $F(1, 23) = 0.131$, $p=0.721$, or condition, $F(1, 23) = 0.013$, $p=0.910$. There was also no significant interaction, $F(1, 23) = 3.004$, $p=0.096$. These results show that the control transition phase of the handover drive ($M = 0.0056$, $SD = 0.0076$) has significantly higher variance in horizontal gaze distribution than the baseline drive ($M = 0.00312$, $SD = 0.00227$).

For vertical gaze distribution, a comparison between baseline and handover drives was made. This test resulted in a significant effect of drive $F(1, 21) = 5.508$, $p=.029$. There

was no significant effect of condition $F(3, 21) = 5.99$, p=.623 or an interaction effect, $F(3, 21) = 0.907$, p=0.454. These results indicate that there is a significant difference in vertical gaze distribution between the baseline (M=0.00739, SD = 0.0057) and handover drives (M = 0.0049, SD=0.00374).

A 2 (Drive) x 2 (Display presence) split-plot ANOVA was conducted to determine if there were any differences in vertical gaze distribution between the display and control (no display) conditions and the drives. This analysis yielded no significant differences between drive, $F(1, 23) = 2.237$, p=.148, or condition $F(1, 23) < 0.001$, $p = 0.986$. There was also no significant interaction of drive and display presence $F(1, 23) = 1.883$, p=.183.

3.4.3 Hypothesis 2c

The representational display will enhance SA more than quantitative and qualitative displays.

This hypothesis was not evaluated due to the findings of Hypothesis 2b that display presence alone did not make a difference in subjective situation awareness.

3.4.4 Hypothesis 2d

During the handover drive, in the display conditions, there will be decreased variance in following distance when compared to the control (no display) condition.

There was no significant difference found between display conditions for variance in following distance $F(3, 50) = 0.907$, $p = .444$. All participants had very high variance in following distance (M = 714,000, SD = 170,000). This is likely due to the difficulty of the

task. At times the lead car would accelerate around a corner that typically, human drivers would slow down for. As participants were instructed to drive as safely as possible, this led to conflicting priorities at times throughout the drives.

3.4.5 Situation awareness summary

These results suggest primarily a change in subjective situation awareness experienced by the participants. There was a difference between the baseline and control transition phase of the handover drive in terms of vertical gaze distribution. This pattern of primarily subjective differences follows the workload measures pattern of results. The gaze distribution measure of situation awareness was not found to be correlated with the SART scores. This indicates that gaze distribution may not be capturing the same level of situation awareness as SART. SART items are at much higher levels of cognitive processing than simple attention allocation as measured by gaze distribution.

3.5 Driving performance

3.5.1 Hypothesis 3a

During the handover drive, in the display conditions, there will be less lane deviation than in the control condition.

A one-way between subjects' ANOVA was used to determine if there were differences in lane deviation during the manual phase of the handover drive due to display type. The results of the test show that there were no significant differences between the display conditions, $F(3, 50) = .414$, $p=0.743$. Therefore, the displays did not impact the

participants' ability to keep the vehicle near the center of the lane after the automation failed.

An independent t-test was done to further test of this analysis was done to determine if there were differences between the control (no display) and display conditions. There were no significant differences in lane deviation during the manual driving phase of the handover drive between the display and control groups, $t(15.923) = 0.410$, $p=.687$.

3.5.2 Hypothesis 3b

Representational displays will result in decreased lane deviation during the handover drive than the other displays.

This hypothesis was not tested as there was no significant difference between conditions in the omnibus ANOVA (Hypothesis 3a).

3.5.3 Driving performance summary

There were no significant differences in lane deviation seen across display conditions in the handover drive. This null result shows that the addition of the automation reliability display did not impact participants' ability to maintain lane position while driving manually.

3.6 Trust

3.6.1 Hypothesis 4a

4a: Trust will be higher in the baseline drive, with no failures, than the handover drive in which failures occur.

Prior to testing the hypothesis an evaluation of how related ratings were the negatively (distrust) and positively (trust) worded items in the Trust in Automation scale. A correlational analysis showed that the baseline trust and distrust subscales were significantly negatively correlated, r=-0.399, p=.001. The same is true of the handover drive trust and distrust subscales, r=-0.543, p<0.001. Although these subscales were moderately correlated, the analysis on each was done independently as they were not highly correlated suggesting that they may capture different variance and add depth to the discussion.

The scores from the Trust in Automation scale were used as two subscales rather than as a single number to determine if there were differences between how trust (positively worded items) and distrust (negatively worded items) were rated across drives. A 2 (Drive) x 4 (Display Type) mixed ANOVA showed a significant difference between the trust scores of baseline and handover drives $F(1, 58) = 34.252$, p<.001. A significant interaction between drive and condition was also found $F(3, 58) = 2.928$, p<.041. There was no significant effect of display type, $F(3, 58) = 0.402$, p=.752. A graphical representation of the interaction can be seen in Figure 14.

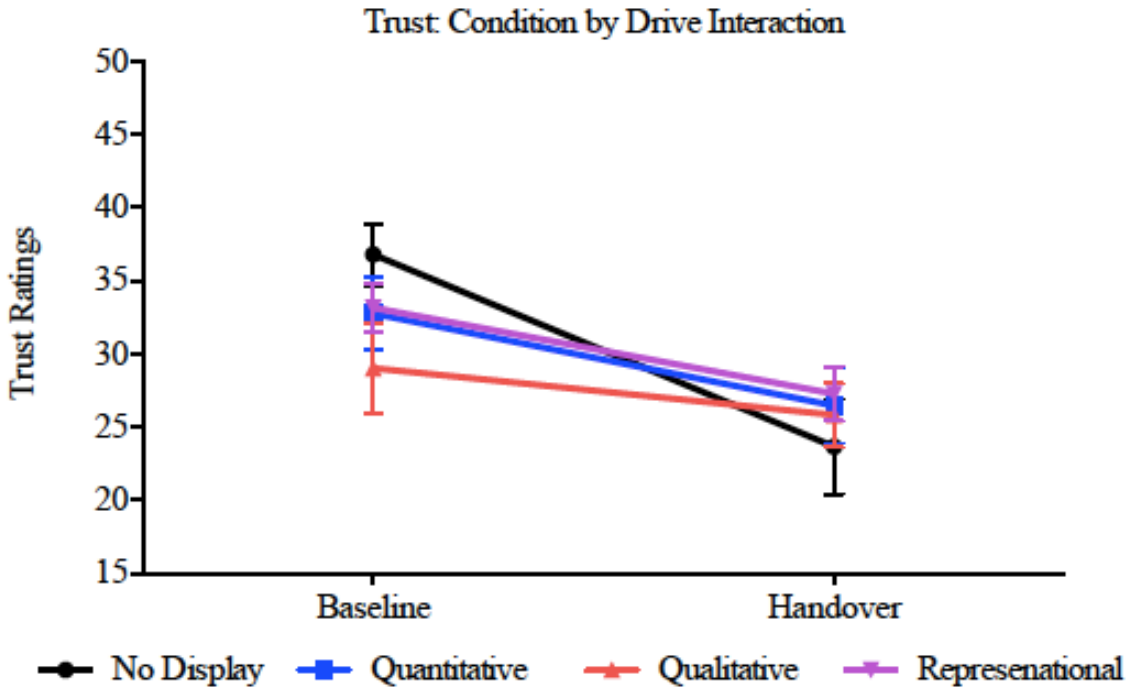


Figure 14. Comparison of trust ratings by condition for the baseline and handover drives.

Since there was no main effect of condition found in the omnibus ANOVA, the focus of the interaction decomposition was on the comparison between baseline and handover drive controlling for display type. A Bonferroni corrected alpha level of p=.0125 was used to protect for Type 1 error. Paired t-tests controlling for display condition and to compare between baseline and handover drive were completed. For the no display condition, there was a significant difference between baseline (M= 36.79, SD = 7.758) and handover drive (M = 23.64, SD = 12.119) assessments of trust $t(13) = 4.490$, p=.001. A significant difference between the baseline (M = 32.75, SD = 9.685) and handover drives (M = 26.44, SD = 10.269) of the quantitative display group was also found $t(15) = 4.056$, p=.001. No significant difference was found between baseline (M = 29.00, SD = 12.928) and handover drives (M = 25.82, SD = 9.002) within the qualitative display condition $t(16)$

= 1.030, p=.318. Finally, the comparison between drives for the representational display group was also found to be statistically different $t(14) = 3.543$, p=.003 with baseline (M = 33.13, SD = 6.198) trust reported higher than the handover drive (M = 27.27, SD = 7.176).

The same pattern of results was found for the distrust subscale. There was a significant difference between the baseline and handover drives, $F(1, 58) = 34.404$, p<.001, as well as a significant interaction of display type and drive $F(3, 58) = 2.816$, p<.047. There was no significant effect of display type, $F(3, 58) = 0.030$, p=.993. A graphical representation of the interaction can be seen in Figure 15.

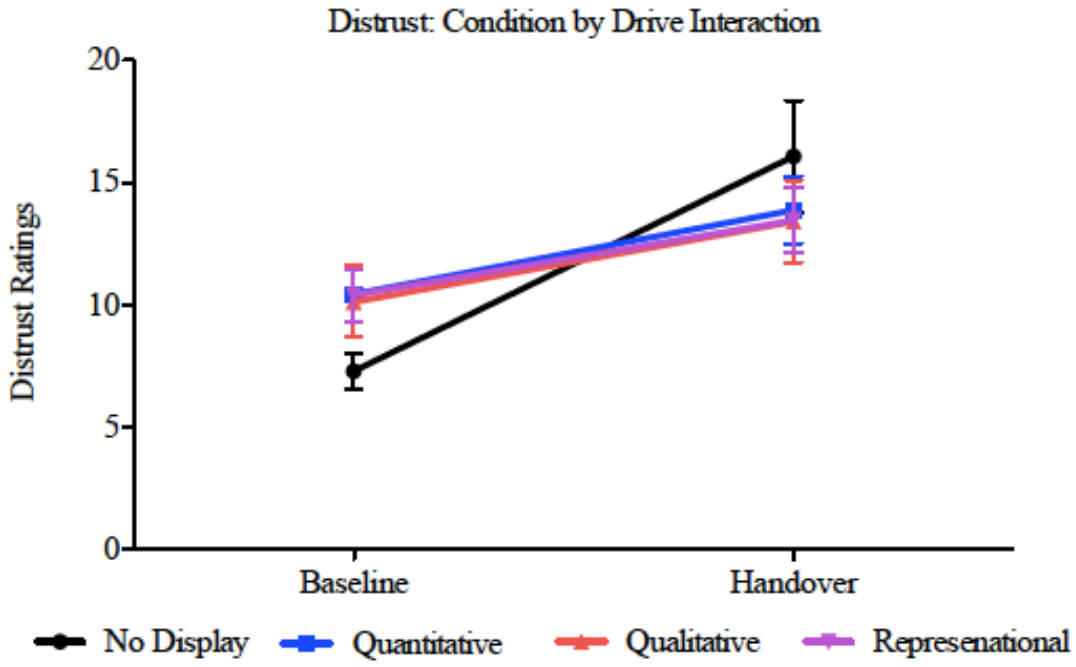


Figure 15. Comparison of distrust ratings by condition for the baseline and handover drives.

The interaction was decomposed using the same methodology as the trust subscale with the same Bonferroni corrected alpha level of p=.0125. For the no display condition, a

significant difference between baseline (M = 7.29, SD = 2.730) and handover drives (M = 16.07, SD = 8.553) was found $t(13) = 3.755$, p=.002. No significant difference was found in the quantitative display condition, $t(15) = 2.391$, p=.300, between the baseline (M = 10.44, SD = 4.718) and handover drives (M = 13.88, SD = 5.500). Similarly, for the qualitative display condition, there was no significant difference found between the baseline (M = 10.12, SD = 6.030) and handover drives (M = 13.41, SD = 6.965), $t(16) = 2.293$, p=.043. Representational displays did have significant differences between baseline (M = 10.40, SD = 4.222) and handover drive (M = 13.47, SD = 5.125), $t(14) = 3.460$, p<.004.

In addition to the Trust in Automation scale, participants were asked to rate the degree to which they agree or disagree with the following statement: "If asked to drive again, I would use the automated lane keeping system." This question was asked only once after the handover drive. There were four options presented ranging from strongly agree to strongly disagree, forcing participants to choose between agreeing to use the lane keeping system again or not. A one-way ANOVA of the responses yielded no significant results between conditions, $F(3, 53) = 0.465$, p=.708. Therefore, the choice of using the automation again in a theoretical future drive was not impacted by the display condition.

3.6.2 Hypothesis 4b

During the handover drive, trust will be higher in the display conditions than in the control (no display) condition.

A 2 (Drive) x 2 (Display presence) mixed ANOVA was completed to test the effects of displays on trust. There was a significant main effect of drive, $F(1, 60) = 39.940$, p<.001,

showing that there was a significant difference between the baseline and handover drive subjective trust ratings. There was also a significant interaction, $F(1, 60) = 7.868$, p=0.007, suggesting that the combination of drive and condition effects trust ratings. There was no significant difference between having a display or not, $F(1, 60) = 0.224$, p=.637.

Likewise, a 2 (Drive) x 2 (Display presence) mixed ANOVA was completed to test the effects of displays on distrust. Similar to the trust results, there was a significant difference between the baseline and handover drive ratings of distrust, $F(1, 60) = 41.610$, p<.001. There was also a significant interaction of drive and condition $F(1, 60) = 8.706$, p = .005. There were no significant differences between conditions for the distrust ratings, $F(1, 60) = 0.035$, p=.852.

An independent samples t-test was completed to determine if there was a significant difference between the control and display conditions in willingness to use automated lane keeping again. The results showed that there was no difference between the control and display groups, $t(11.699) = 0.723$, p=.484.

These findings support those of Hypothesis 4b showing that the combination of the display and the drive makes a difference in trust and distrust ratings. However, the presence of the display alone does not impact feelings of trust and distrust.

3.6.3 Trust summary

For both trust and distrust subscales of the Trust in Automation Scale, there were significant differences between the baseline and handover drives. For the trust subscale, ratings were higher for the baseline drive than the handover drive meaning that the

participants trusted the system more during the first drive than the second. For the distrust subscale, ratings were higher for the handover drive than the baseline drive. This means that participants had greater feelings of distrust toward the system after the handover drive than after the baseline drive. Together, these results mean that the participants had higher levels of trust and lower levels of distrust after completing the baseline drive than they reported after experiencing the handover drive. This shows that the automation failure caused a decrease in trust and an increase in distrust of the automated lane keeping system.

Not all participants experienced the increase in distrust and decrease in trust as significantly as others, as evidence by the interaction effects. The no display, quantitative, and representational groups had significantly lower trust ratings after the handover drive than after the baseline drive. Therefore, the qualitative display reduced the impact of automation failure on trust whereas the other displays did not mitigate this effect significantly. The no display and representational groups had significantly higher distrust levels for the handover drive than the baseline drive. Therefore, the quantitative and qualitative displays reduced the impact of automation failure on distrust feelings.

3.7 Other results

3.7.1 Handover experience questionnaire

The handover experience questionnaire (APPENDIX I. Handover Experience Questionnaire) was administered to participants after completing the handover drive. Higher scores on this measure indicate higher levels of confidence during the handover and predictability of the failure prior to it occurring. A one-way ANOVA of the responses was completed to determine if there were any differences between display conditions. There

was no main effect of display condition $F(3, 57) = 2.608$, p=.060. To follow this analysis, a t-test comparing the no display (control) condition to the display conditions was completed. There was a significant difference between these two groups, $t(18.13) = 2.706$ p=.014, with the no display condition (M= 27.917, SD = 5.794) having significantly lower handover experience scores than the display conditions (M=23.077, SD = 5.433). This indicates that participants with reliability displays were more confident in their abilities to take over control of the vehicle and felt that they knew the failure was going to occur prior to it occurring. This is further evidence for different subjective experiences due to the displays despite lack of objective measure support.

3.7.2 Total time spent looking at the display screen

Time spent looking at the display screen (where the reliability display was located) was calculated by drive summing the total gaze time for each drive. This was then converted to a percentage for comparison across baseline and handover drive since the handover drive is over twice as long as the baseline drive. No formal analyses were completed on this set of data as there were many technical difficulties that led to very small sample sizes per group. The breakdown of the percentage of time spent looking at the display screen for each group can be seen in Table 6. The overall small percentage of time spent looking at the displays indicates that participants spent much more time attending to the roadway and mirrors than they did attending to the reliability display. The numerical values seem to indicate that the qualitative display participants spent more time looking at the display than the other groups.

	Condition	**Mean percentage**	**Standard deviation**

Baseline drive	No display	0.195	.346
	Quantitative	0	0
	Qualitative	1.191	1.067
	Representational	0.866	1.341
Handover drive	No display	0.123	0.261
	Quantitative	0.285	0.403
	Qualitative	1.626	1.274
	Representational	0.819	1.039

Table 6. Mean percentage gaze time spent on display screen by drive and condition.

CHAPTER 4. DISCUSSION

The findings of this research suggest that there are subjective differences between driving with automation (baseline drive) and driving with automation that fails (handover drive). This difference was seen in the subjective measurement of workload, situation awareness, and trust. This research also confirmed that the presence of an automation reliability display of any information type impacts trust when compared to no display. Specifically, quantitative and qualitative displays are most effective in aiding trust calibration.

Many measures show differences between the baseline drive (without an automation failure) and the handover drive (with an automation failure) which help characterize the differences between the tasks. Subjective workload, as measured by weighted NASA-TLX, was higher in the handover drive than the baseline drive. Although this was not echoed in the objective measures, this subjective experience of higher workload was hypothesized and expected from experiencing automation failure. This difference in subjective experience is also echoed in the SART ratings. Participants rated themselves to be much more situationally aware in the baseline drive than in the handover drive. This is likely due to the unexpected automation failure that they experienced. Finally, trust and distrust differences were found between both drive and display type which indicates that there were differences due to automation failure and display type on these constructs.

4.1 Automated driving task characterization

Characterization of the dynamic task of driving with automated systems has been accomplished by collecting heart rate, pupil diameter, and gaze distribution measures throughout each drive. The heart rate data show that throughout the baseline drive, there is an overall decrease in workload as drivers become more familiar with the task environment and is sensitive enough to detect changes in workload due to increases in road curvature. In the handover drive, increased heart rate is seen due to the reduction in reliability, followed by further increase during control transition, and finally a general downward trend as manual driving task familiarity increases.

Pupil diameter, also a workload proxy, tells a similar story to heart rate. The baseline drive is characterized by decreasing pupil diameter as familiarity with the task increases. In the test drive, as reliability decreases, pupil size increases leading up to the control transition. At the point of control transition there is a further increase in pupil diameter which then leads to a general decreasing trend as the manual driving task becomes more familiar.

The gaze distribution data suggest that during the baseline drive there is a slightly increasing area of attention as familiarity increases and workload decreases. This suggests a lack of perceptual tunneling. The handover drive, however, begins similarly to the baseline drive with a slight increase in gaze distribution. As the reliability level decreases and road curvature increases, there is an increase in gaze distribution. As the control transition occurs, there is a sharp decline in gaze distribution indicating that perceptual tunneling may be occurring.

This dynamic characterization of the task gives insight into how drivers react to different roadway and automation changes as they navigate a driving route. This characterization can lead to displays and warning systems that are timed appropriately with the changing vehicle and roadway dynamics to avoid overload or allow for optimally timing secondary task interruption at higher levels of automation.

4.2 Evidence for trust calibration

Findings from the trust and distrust data show that more calibration occurred for participants in the quantitative and qualitative display conditions than in the no display condition. Drivers of automated vehicles with reliability displays will be able to appropriately calibrate their trust to system performance, reducing the impacts of automation failure on their feelings toward the system. Further, drivers that do not have access to this information will have much more extreme reactions to the automation failure which could lead to disuse over time. This finding suggests that all vehicles should be equipped with these displays to better inform drivers of system performance throughout the drive and reduce negative consequences of imperfect automation.

4.3 Situation awareness measurement techniques

Gaze distribution, variance of gaze position in the x and y directions, was not found to be correlated with the SART. This could indicate that the two measures are at different levels of situation awareness. SART questions are focused on higher level processes than simply attending to visual information. For example, one question requires participants to retrospectively determine how complex the situation was, “How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and

straightforward (Low)? (1)" (Taylor, 1990). This is a much higher level of processing than simply looking at a given area of the screen.

While visual attention is necessary, in this case, to understand the driving environment and displays, it is only the first step towards situation awareness as described by Endsley's process model (Endsley, 1995b). The types of questions asked by SART would fall in Endsley's levels two (understanding) and three (projection); whereas gaze distribution is the first step towards perception, Endsley's first level (Endsley, 1995b). Correlations of gaze distribution may be higher if the participants were asked questions pertaining to whether they saw a specific stimulus in the driving environment rather than whether they perceived the driving environment to be complex.

4.4 Design implications

The results of this study suggest that any reliability display is much better than no display. Specifically, reliability displays allow drivers to calibrate themselves to the system performance over time reducing the effects of automation failure on their feelings of trust toward the system.

Based on the trust and distrust results, qualitative displays seem to mitigate the effects of automation failure for both trust and distrust measures. Quantitative displays were found to mitigate increases in distrust due to automation failure but not increases decreases in trust. For these reasons, quantitative and qualitative displays, which are direct or slight abstractions of numeric information, have proven to aid in appropriate trust calibration. This type of numeric, or slightly abstracted numeric, display should be

considered for presenting reliability information to drivers such that they can calibrate their trust in the system effectively.

As seen in the gaze distribution data, there was very little variance in gaze position throughout either baseline or handover drive. This indicates that the automated driving task still demanded high levels of visual attention on the roadway. Future visual displays should be more centrally located, potentially through heads up displays, to give drivers' access to that information without having to move their gaze from the central windshield area. Alternatively, other modalities of reliability displays should be explored to determine if displays requiring less or no visual attention would have greater impacts on handover and lead to better trust calibration.

4.5 Limitations

This study was designed with two goals in mind. First to better characterize the experience of driving with automated lane keeping and second, to determine the impacts of reliability displays on that experience. To balance these two goals, there was an extended time of manual driving after the automation failure occurred in the handover drive that led to approximately seven minutes between the automation failure and subjective assessments. This time (and the experience of manual driving) impacted the participants' perspectives when responding to the subjective scales and questionnaires. They could have been reflecting on an average experience of the drive, one specific portion of the drive, or something between those and there is no way of knowing what they were thinking specifically when answering those questions. In the future, there should be less or no

manual driving after failure such that their reflection on the automation failure and handover is coming from a more recent experience.

Due to technical failures, there were very small sample sizes in the eye tracking and heart rate measures. This small sample size likely impacted the results of these measures greatly. Future studies should include higher sample sizes to help mitigate any unforeseen technical issues.

4.6 Future research

Future research should investigate other ways of presenting reliability information, through manipulating the metric of presentation, the location of the visual display, and the modality(ies) of presentation. These manipulations will give greater insight into the results of this study to determine if the location of the display or metric presented in this study impacted the handover experience results. As evidence of the small variance in gaze distribution both horizontally and vertically, future studies should specifically evaluate auditory versions of these displays to determine if reduced visual attention required to the reliability displays impacts workload, situation awareness, and driving performance measures.

Higher levels of automation may allow drivers to attend to visual displays more throughout the drive. This study focused on a very low level of automation where the driver was still highly engaged with the driving task. Moving up to SAE Level 3 automation may greatly change how drivers interact with the displays tested in this experiment. This higher level of automation should lead to less engagement required by drivers and therefore more resources to attend to displays. Future research should investigate whether visual displays

are more impactful at higher levels of automation and how they manipulate handover experience.

APPENDIX A. STUDY RECRUITMENT FORM

Recruitment Description Listing for School of Psychology Subject Pool Website

We are looking for participants to be subjects in a 1.5-hour study session (1.5 credit total). Participants will be asked to drive in a driving simulator. This research will help identify the most effective types of displays. Participants must have normal or corrected to normal vision, mobility, and hearing and have 2 years minimum of driving experience and a valid license.

Recruitment Description Listing for word of mouth postings

We are looking for volunteers to be subjects in a 1.5-hour study session. Participants will be asked to drive in a simulator. This research will help identify the most effective types of dashboard displays. Participants must have normal or corrected to normal vision, mobility, and hearing and have 2 years minimum of driving experience and a valid license.

APPENDIX B. CONSENT FORM

Multimodal Driving Display Study
Consent Form

Study: Comparison of Multimodal Displays for Driving
Principal Investigator: Dr. Bruce N. Walker (404-894-8265)
Location: School of Psychology, Coon (Psychology) Building, Georgia Institute of Technology
Duration of Each Session: 1.5 hours **Number of Sessions:** 1
Total Compensation: 1.5 credit hours (if students)
Approximate Number of Participants: 120
Participation limitations: Normal or corrected to normal vision and hearing and no mobility impairments. Participants must also have had a valid driver's license for 2 years and be wearing contacts or glasses on the day of the study if vision correction is necessary.

General: You are being asked to volunteer for a psychological experiment. Displays have one of the most important roles in the driving task: informing users of the status of the car. We want to understand the effect of displays on driving. You will be asked to perform two driving tasks. We will be taking a variety of measurements, such as eye movement, driving performance, heart rate, your awareness of the road conditions, and task workload. Your participation will help us investigate the effectiveness of different driving display setups. It will also provide you with some experience in the conduct of research in psychology.

Study Purposes: This research is looking at how different types of displays can affect driver performance, safety, and other measures.

Procedures: You will be given a consent form and an instruction form explaining the full procedures if you decide to participate in this study. In summary: You will read through the instruction form and ask the experimenter any questions you have. After you finish reading the instruction form, you will be taken through a simulator sickness screening to make sure the driving simulator will not cause you physical discomforts. When you finish the sickness screening, you will be asked to complete two driving tasks. The experimenters will go over the instructions for each drive. You will be asked to fill out questionnaires throughout the study. At the conclusion of the experiment, you will be debriefed on our study, and released. During the study, you may be led to believe some things that are not true. When the study is over, we will tell you everything. At that time you can decide whether to let us use your information. You have the right to then require that your information be destroyed and not be used in the study.

Foreseeable Risks or Discomforts: This study is expected to involve no more than minimal risks associated with listening to sounds and driving a low-fidelity simulator. A small number of people may feel physically sick when using a driving simulator. In the event that you do experience any sickness at that time or at any point during the study, we

will ask you to sit in a stationary chair until the feeling subsides. At that point you will be debriefed and released from the study but will still receive your full credit for participating in the study.

Confidentiality: The following procedures will be followed to keep your personal information confidential in this study: The data that is collected about you will be kept private to the extent allowed by law. To protect your privacy, your records will be kept under a code number rather than by name. Your records will be kept in locked files and only study staff will be allowed to look at them. Your name and any other fact that might point to you will not appear when results of this study are presented or published. To make sure that this research is being carried out in the proper way, the Georgia Institute of Technology IRB will review study records. The Office of Human Research Protections may also look over study records during required reviews. Again, your privacy will be protected to the extent allowed by law.

Alternative Credit Option: Alternatives to participating in this study are provided by your course instructor. They include, but are not limited to, reading journal articles and writing a brief report based on the articles.

Injury/Adverse Reaction: Reports of injury or reaction should be made to Dr. Bruce Walker (404-894-8265). Neither the Georgia Institute of Technology nor the principal investigator has made provision for payment of costs associated with any injury resulting from participation in this study.

Contact Persons: If you have questions about this research, call or write Dr. Bruce Walker at 404-894-8265; School of Psychology, GA Tech, 654 Cherry Street, Atlanta, GA 30332-0170.

Statement of Rights: You have rights as a research volunteer. Taking part in this study is completely voluntary. If you do not take part, you will have no penalty. You may stop taking part in this study at any time with no penalty. If you have any questions about your rights as a research volunteer, call or write: The Institutional Review Board, Office of Research Integrity Assurance, 505 Tenth Street, Campus 0420. Phone: 404-894-6942; Fax: 404-385-2081.

Benefits to you: This study will contribute to automobile interface design by discovering more effective ways to present information. Other than the potential 1.5 credits you will receive for participating there is no other benefit to you.

Costs to you: There are no costs to you other than your time for participating in this study.

Signatures: A copy of this form will be given to you. If you sign below, it means that you have read the information given in this consent form, and you would like to be a volunteer in this study.

Participant's Signature:
_____

Date: _____

Person Obtaining Consent:
_____

Date: _____

APPENDIX C. DEMOGRAPHICS QUESTIONNAIRE

1. Do you have a driver's license?
    a. Yes (If yes, "How many years have you held a drivers' license?")
    b. No (if no, the following is displayed: "Sorry, you are not eligible for participation in this study. Please speak to the researcher.")
2. What is your age?
3. What is your gender?
    a. Male
    b. Female
    c. Choose not to identify
4. What is your primary language?
5. What other languages do you speak?
6. How many hours per week do you drive when you are on campus?
7. How many hours per week do you drive when you are not on campus?
8. What is your level of familiarity with automated safety features such as automated lane keeping? Automated lane keeping systems automatically steer the vehicle to maintain position within a lane.
    a. I own a vehicle with one or more automated safety features.
    b. I have driven a vehicle with one or more automated safety features.
    c. I have ridden in a vehicle with one or more automated safety features.
    d. I am familiar with automated safety features.
    e. I have never heard of automated safety features prior to this study.

APPENDIX D. PARTICIPANT INSTRUCTIONS

Thanks and Introduction

First of all, thank you for your participation in this study. We are members of Sonification Lab in school of psychology.

Purpose of Experiment

This research is investigating the effect of displays on driving.

Procedure

Consent

The consent form presented to you is to inform you of the content of this experiment. Please read through it, and ask any questions you have before you sign it. During the experiment, please let us know if you have questions, concerns, discomforts, or would like to withdraw from the experiment. You can do so without penalty.

General Instructions

Before this experiment, we will ask you to complete a simulator sickness screening first. This is to ensure that you do not encounter any motion sickness during the experiment. Then you will be asked to complete a set of questionnaires. Next, you will complete the first of two driving scenarios which will be followed by another set of questionnaires. The second drive will follow with a final set of questionnaires. The session should last no longer than an hour and a half, and the experimenter(s) will help you throughout the session.

Sim Sickness Screening

To make sure the driving simulator will not cause you any physical discomfort, we will conduct a screening procedure. This procedure includes a pre-drive survey, a short drive, and a post-drive survey. If for any reason, you feel sick during the procedure, this session will end and you will receive full credit for your time here.

Questionnaires and Driving Tasks

We will be collecting data during two separate driving tasks. The first driving course will be about 10 minutes and the second driving course will be about 20 minutes in length. Each drive will have its own set of instructions that the experimenters will go over with you before the drive. We will be collecting a number of measurements, such as eye movement, heart rate, driving performance, and workload. The driving scenario will be similar for both of the driving tasks.

Debrief

Once the final set of questionnaires is completed after the second drive, you will be debriefed on the experiment and released. We will then assign you credit for your participation.

APPENDIX E. PARTICIPANT DRIVING INSTRUCTIONS

Driving Task #1

The driving course will last about 6 to 8 minutes. For this drive, we would like you to follow the blue car ahead of you at the same distance that you see at the beginning of the drive throughout the duration of the drive (about 50 ft). During this drive, you will be using automated lane keeping. Automated lane keeping systems keep the vehicle in the center of the lane so that you do not have to steer. Putting your hands on the steering wheel and turning it even a small amount can turn off the automated lane keeping system. To avoid doing so accidentally, please keep your hands off the steering wheel unless you feel that you must take control of the vehicle to avoid an accident. You will know that the automated lane keeping system is on if the green, nearly parallel lines are present in the dashboard. If this display is not present, the automated lane keeping system is not on.

- For participants in the quantitative display condition:
    - In addition to the display in the dashboard that tells you whether the automated lane keeping system is on or off, located on the side screen, you will also have a display showing a percentage throughout the drive. This percentage represents the reliability of the system throughout the drive. A high percentage indicates high reliability *[show them where the display is on the screen]*. High reliability means that the system is able to perform the automated lane keeping task well. *[Flip to the moderate reliability display and show it to them then flip to the low reliability display]* A low percentage indicates low reliability. Low reliability indicates that the system is unable to perform the automated lane keeping task well *[make sure to point out the low reliability display]*. The display will update throughout the drive to inform you of the system performance over time.
- For participants in the qualitative display condition:
    - In addition to the display in the dashboard that tells you whether the automated lane keeping system is on or off, located on the side screen, you will also have a display showing a vertical bar that empties and fills as reliability changes. A full bar indicates high reliability *[show them where the display is on the screen]*. High reliability indicates that the system is able to perform the automated lane keeping task well. *[Flip to the moderate reliability display and show it to them then flip to the low reliability display]* The less full the bar appears, the lower the reliability the system has. Low reliability indicates that the system is unable to perform the automated lane keeping task well *[make sure to point out the low reliability display]*. The display will update throughout the drive to inform you of the system performance over time.
- For participants in the representative display condition:
    - In addition to the display in the dashboard that tells you whether the automated lane keeping system is on or off, located on the side of the screen, you will also have a display showing a vertical road. A static road indicates that the system has high reliability. *[Show them the high*

display] High reliability indicates that the system is able to perform the automated lane keeping task well. A road with an exclamation mark within a triangle represents moderate reliability. *[Show them the moderate display]* A vertical road that moves with an exclamation mark inside of a triangle indicates that the system has low reliability. Low reliability indicates that the system is unable to perform the automated lane keeping task well *[Show them the low display].* The display will update throughout the drive to inform you of the system performance over time.

- Participants in the no display condition will not be given any additional instructions.

If you see any speed limit signs throughout the drive, ignore them. Your primary goal is to maintain 50 feet of distance between you and the blue car. Please complete the drive as safely as possible. Do you have any questions before we get started?

Driving Task #2

The driving course is similar to the last drive you just completed and will last about 15 minutes. For this drive, we would again like you to follow the blue car ahead of you at the same distance that you see at the beginning of the drive throughout the duration of the drive (50 ft). You will also be using the automated lane keeping system in this drive. Automated lane keeping systems keep the vehicle in the center of the lane so that you do not have to steer. Remember, putting your hands on the steering wheel and turning it even a small amount can turn off the automated lane keeping system. To avoid doing so accidentally, please keep your hands off the steering wheel unless you feel that you must take control of the vehicle to avoid an accident.

If the participant is in the display condition:

- You will also have the same display on the side screen that you had in the previous drive.

If you see any speed limit signs throughout the drive, ignore them. Your primary goal is to maintain 50 ft of distance between you and the blue car. Please complete the drive as safely as possible. Do you have any questions?

Completion

Upon completion, we will stop the driving scenario and present you with the final set of questionnaires to complete.

APPENDIX F. SITUATIONAL AWARENESS RATING TECHNIQUE (SART)

Based on the drive you just completed, please fill out the following questions:

Instability of Situation

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
How changable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)? (1)	❍	❍	❍	❍	❍	❍	❍

Complexity of Situation

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)? (1)	❍	❍	❍	❍	❍	❍	❍

Variability of Situation

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)? (1)	❍	❍	❍	❍	❍	❍	❍

Arousal

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)? (1)	❍	❍	❍	❍	❍	❍	❍

Concentration of Attention

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)? (1)	❍	❍	❍	❍	❍	❍	❍

Division of Attention

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)? (1)	❍	❍	❍	❍	❍	❍	❍

Spare Mental Capacity

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
How much mental capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)? (1)	❍	❍	❍	❍	❍	❍	❍

Information Quantity

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
How much information have you gained about the situation? Have you received or understood a great deal of knowledge (High) or very little (Low)? (1)	❍	❍	❍	❍	❍	❍	❍

Familiarity with Situation

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation(Low)? (1)	❍	❍	❍	❍	❍	❍	❍

APPENDIX G. TRUST IN AUTOMATION SCALE

Based on the drive you just completed, please fill out the following questions where 1 = not at all and 7 = extremely.

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
The system is deceptive. (1)	❍	❍	❍	❍	❍	❍	❍
The system behaves in an underhanded manner. (2)	❍	❍	❍	❍	❍	❍	❍
I am suspicious of the system's intent, action, or outputs. (3)	❍	❍	❍	❍	❍	❍	❍
I am wary of the system. (4)	❍	❍	❍	❍	❍	❍	❍
The system's actions will have a harmful or injurious outcome. (5)	❍	❍	❍	❍	❍	❍	❍
I am confident in the system. (6)	❍	❍	❍	❍	❍	❍	❍
The system provides security. (7)	❍	❍	❍	❍	❍	❍	❍
The system has integrity. (8)	❍	❍	❍	❍	❍	❍	❍
The system is dependable. (9)	❍	❍	❍	❍	❍	❍	❍
The system is reliable. (10)	❍	❍	❍	❍	❍	❍	❍
I can trust the system. (11)	❍	❍	❍	❍	❍	❍	❍
I am familiar with the system. (12)	❍	❍	❍	❍	❍	❍	❍

APPENDIX H. NASA-TLX SCALE DEFINITIONS

Title	Endpoints	Description
Mental Demand	Low/High	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical demand	Low/High	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal demand	Low/High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Good/Poor	How successful do you think you were accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration level	Low/High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

APPENDIX I. HEART RATE MONITOR PLACEMENT

INSTRUCTIONS

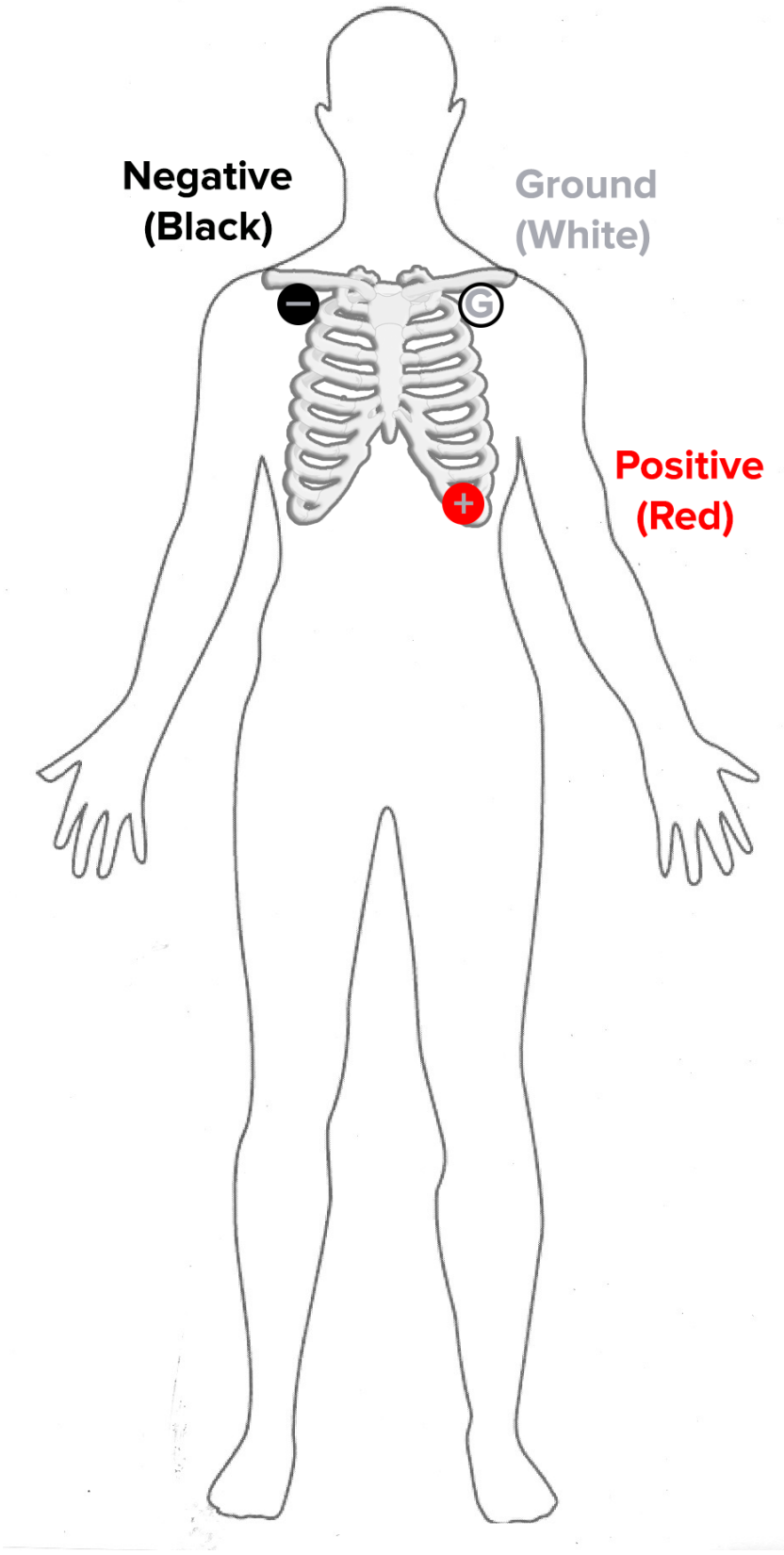
Instructions for Applying Heart Rate Monitoring Leads

1) Using the included diagram, identify the 3 locations where the electrode pads will be applied. Two will be located just under the collarbones, preferably in the gap between the shoulder muscles. The other will be near the stomach over the bottom rib bone.
2) Gently wipe the 3 areas with a cotton swab to clear any dead skin.
3) Using an alcohol wipe, clean the 3 areas thoroughly and then let them dry to allow the electrode pads to stick cleanly.
4) Once the wiped areas are dry, place one pad in the center of each cleaned area. All pads are the same, and it does not matter which pads goes on which of the 3 areas.
5) Grab the three wire leads for the heart rate system. They are labeled with twist ties as:

+ - G

(Positive) (Negative) (Ground)

6) Snap each lead onto the proper electrode pad according to the included diagram. The leads should snap in with only a small amount of force.
7) Verify the location of the three electrodes. They should be free from all fabric, belts, and clothing and should not fall or peel off as you move. The leads should go under your clothing and out of the area by your belt.
8) Notify the experimenter that you are ready to continue.


Negative
(Black)
Ground
(White)
Positive
(Red)
G

APPENDIX I. HANDOVER EXPERIENCE QUESTIONNAIRE

Please answer the following questions in relation to the drive you just completed.

Anchors: Strongly Disagree (1) and Strongly Agree (7)

1. I knew that the system was going to fail prior to the failure occurring.
2. I felt prepared to take control of steering the vehicle.
3. I felt safe during the transition from automated to manual steering.
4. I felt confident in my abilities to begin steering the vehicle.
5. I felt confident in the system's abilities.
6. I became more aware of the system's ability to keep my car in the lane throughout the drive.

APPENDIX J. DEBRIEF

Thanks and Introduction
First of all, thank you for your participation in this experiment. We are members of Sonification Lab in the School of Psychology.

Purpose of Experiment
The purpose of this experiment was to investigate the effects of automation reliability displays on driving performance, workload, your awareness of the driving environment, and driving performance. In each of the two drives, we measured your eye movements, pupil size, driving performance, workload, awareness of the driving environment, trust in the automated system, and feelings toward the automated system. There were four possible conditions: no display, quantitative display, qualitative display, and representational display. You were randomly assigned to one of these conditions. We did not tell you prior to the second drive that the automation was going to fail. We did this so that we could capture how you reacted to the situation based on the displays and other cues in the driving environment. If you would like to withdraw your consent and data now that you know about this concealment please let the experimenter know at this time and they will ensure all data collected from you is destroyed and not included in the analysis.

Meaning of Expected Results
We expect that participants that received information from the displays about the reliability of the system were more ready to take control of the vehicle during the second drive. We expect that analysis of eye movements, driving performance, workload, awareness of the driving environment, and trust will show advantages to having these displays over no display at all. In addition, we expect that representational displays will prepare drivers for the takeover more than the quantitative or qualitative displays. These results will be used to establish guidelines for the design of displays for automated driving.

Confidentiality and Anonymity
The results of your experiment will be used for only psychological study and never used for any other purposes. The data that is collected from you will be kept private to the extent allowed by law. To protect your privacy, your records will be kept under a code number rather than by name. Your records will be kept in locked files and only research staffs will be allowed to look at them. Your name and any other fact that might point to you will not appear when results of this study are presented or published. To make sure that this research is being carried out in the proper way, the Georgia Institute of Technology IRB will review study records. Again, your privacy will be protected to the extent allowed by law.

Conclusion
All of the experiment procedures are finished. We very much appreciate your efforts again.

Contact Information
For further information of this research, contact:
Principal Investigator
Dr. Bruce Walker (bruce.walker@psych.gatech.edu)
Experimenters
Brittany Noah (brittany.noah@gatech.edu)
Zoe Becerra (zbecerra3@gatech.edu)
Grace Li (tli.grace@gatech.edu)
Mengyao Li (mli@agnesscott.edu)

APPENDIX K. SAE LEVELS OF AUTOMATION

Summary of Levels of Driving Automation for On-Road Vehicles

This table summarizes SAE International's levels of *driving* automation for on-road vehicles. Information Report J3016 provides full definitions for these levels and for the italicized terms used therein. The levels are descriptive rather than normative and technical rather than legal. Elements indicate minimum rather than maximum capabilities for each level. "System" refers to the driver assistance system, combination of driver assistance systems, or *automated driving system*, as appropriate.

The table also shows how SAE's levels definitively correspond to those developed by the Germany Federal Highway Research Institute (BASt) and approximately correspond to those described by the US National Highway Traffic Safety Administration (NHTSA) in its "Preliminary Statement of Policy Concerning Automated Vehicles" of May 30, 2013.

Level	Name	Narrative definition	Execution of steering and acceleration/ deceleration	Monitoring of driving environment	Fallback performance of *dynamic driving task*	System capability (*driving modes*)	BASt level	NHTSA level
***Human driver* monitors the driving environment**								
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a	Driver only	0
1	**Driver Assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes	Assisted	1
2	**Partial Automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes	Partially automated	2
***Automated driving system* ("system") monitors the driving environment**								
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the *dynamic driving task* with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes	Highly automated	3
4	**High Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes	Fully automated	3/4
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**	-	

cyberlaw.stanford.edu/loda

REFERENCES

Bartels, M., & Marshall, S. (2012). Measuring cognitive workload across different eye tracking hardware platforms. *Proceedings of the Symposium on Eye Tracking ...*, 161–164. https://doi.org/10.1145/2168556.2168582

Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the Driver-Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors*, *55*(6). https://doi.org/10.1177/0018720813482327

Bennett, K. B., Nagy, A. L., & Flach, J. M. (2012). Visual Displays. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (4th ed., pp. 1191–1221). Hoboken, NJ: John Wiley & Sons, Inc.

Billings, C. E. (1991). *Human-Centered Aircraft Automation : A Concept and Guideline*.

Dehais, F., Causse, M., & Tremblay, S. (2011). Mitigation of Conflicts with Automation: Use of Cognitive Countermeasures. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 448–460. https://doi.org/10.1177/0018720811418635

Durso, F. T., & Gronlund, S. D. (1999). Situation Awareness. In M. T. H. Durso, F. T., Nickerson, R. S., Schvaneveldt, S. T., Lindsay, D. S., Chi (Ed.), *Handbook of Applied Cognition* (pp. 283–314). Hoboken, NJ: John Wiley & Sons, Inc.

Durso, F. T., Stearman, E. J., Morrow, D. G., Mosier, K. L., Fischer, U., Pop, V. L., & Feigh, K. M. (2015). Exploring relationships of human-automation interaction

consequences on pilots: uncovering subsystems. *Human Factors*, *57*(3), 397–406. https://doi.org/10.1177/0018720814552296

Ellis, K. K. E. (2009). Eye tracking metrics for workload estimation in flight deck operations. *ProQuest Dissertations and Theses*, *1467693*, 115. Retrieved from http://ezproxy.net.ucf.edu/login?url=http://search.proquest.com/docview/304901025?accountid=10003%5Cnhttp://sfx.fcla.edu/ucf?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+&+theses&sid=ProQ:ProQuest+Dissertations+&+T

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National*, 789–795. https://doi.org/10.1109/NAECON.1988.195097

Endsley, M. R. (1995a). Measurement of Situation Awareness in Dynamic-Systems. *Human Factors*. https://doi.org/Doi 10.1518/001872095779049499

Endsley, M. R. (1995b). Measurement of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 65–84. https://doi.org/10.1518/001872095779049499

Endsley, M. R. (1995c). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Endsley, M. R. (1998). Design and Evaluation for situation awareness enhancement. In *Proceedings of the 32nd Human Factors and Ergonomics Society Meeting* (pp. 97–

101).

Endsley, M. R. (2011). Principles of Designing for SA. *Designing for Situation Awareness: An Approach to User-Centered Design*, 83–219. https://doi.org/doi:10.1201/b11371-9

Gable, T. M., & Walker, B. N. (2013). *Georgia Tech Simulator Sickness Screening Protocol*. Atlanta.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*(4), 457–461.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*(C), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Heldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomoys driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '13* (pp. 210–217). ACM.

Ikuma, L. H., Harvey, C., Taylor, C. F., & Handal, C. (2014). A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. *Journal of Loss Prevention in the Process Industries*, *32*, 454–465. https://doi.org/10.1016/j.jlp.2014.11.001

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental

workload in human-computer interaction. In *Chi'04 extended abstracts on human factors in computing systems* (pp. 1477–1489). ACM.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated System. *International Journal of Cognitive Ergonomics*, *4*(1), 53. https://doi.org/10.1207/S15327566IJCE0401_04

Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space, & Environmental Medicine*, *67*(6), 507–612.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113–153. https://doi.org/10.1080/1463922021000054335

Kieras, D. E., & Meyer, D. E. (1995). *An Overview of the EPIC Architecture for Cognition and Performance with Application to Human-Computer Interaction*.

Kun, A. L., Palinko, O., & Razumenic, I. (2012). Exploring the effects of size and luminance of visual targets on the pupillary light reflex. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 183–186). ACM.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(2), 241–256. https://doi.org/10.1518/001872006777724408

Maier, A. M., Baltsen, N., Christoffersen, H., & Strrle, H. (2014). Towards Diagram Understanding: A Pilot-Study Measuring Cognitive Workload Through Eye-Tracking. *Proc. Intl. Conf. Human Behavior in Design*, (October), 1–6.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734.

Mccall, R., McGee, F., Meschtscherjakov, A., Louveton, N., & Engel, T. (2016). Towards A Taxonomy of Autonomous Vehicle Handover Situations. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '16), October 24–26, 2016, Ann Arbor, MI, USA*, 193–200. https://doi.org/10.1145/3003715.3005456

Mehler, B., Reimer, B., Coughlin, J. F., & Dusek, J. a. (2009). Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. *Transportation Research Record: Journal of the Transportation Research Board*, (2138), 6–12. https://doi.org/10.3141/2138-02

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2014). Are Well-Calibrated Users Effective Users? Associations Between Calibration of Trust and Performance on an Automation-Aided Task. *Human Factors*, *57*(1), 34–47. https://doi.org/10.1177/0018720814561675

Noah, B. E., Gable, T. M., Chen, S.-Y., Singh, S., & Walker, B. N. (2017). Development and Preliminary Evaluation of Reliability Displays for Automated Lane Keeping. In *Proceedings of the International Conference on Automotive User Interfaces and*

Interactive Vehicular Applications - AutomotiveUI '17 (pp. 202–208). ACM. https://doi.org/10.1145/3122986.3123007

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 141–144. https://doi.org/10.1145/1743666.1743701

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced "Complacency." *The International Journal of Aviation Psychology*, *3*(1), 1–23. https://doi.org/10.1207/s15327108ijap0301_1

Parasuraman, R., & Riley, V. (1997). Humans and automation : Use , misuse , disuse , abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans : A Publication of the IEEE Systems, Man, and Cybernetics Society*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology. Applied*, *9*(2), 119–37. https://doi.org/10.1037/1076-898X.9.2.119

Sanders, M. S., & McCormick, E. J. (1993). *Human Factors in Engineering and Design* (7th ed.). New York, NY: McGraw Hill.

Sarter, N. B., & Woods, D. D. (1995). How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 5–19. https://doi.org/10.1518/001872095779049516

Sarter, N. B., & Woods, D. D. (2000). Team play with a powerful and independent agent: A full-mission simulation study. *Human Factors2*, *42*(3), 390–402.

Seppelt, B. D., & Lee, J. D. (2007). Making adaptive cruise control (ACC) limits visible. *International Journal of Human Computer Studies*, *65*(3), 192–205. https://doi.org/10.1016/j.ijhcs.2006.10.001

Sheridan, T. B. (2012). Human Supervisory Control. In G. Slavendly (Ed.), *Handbook of Human Factors and Ergonomics* (4th ed., pp. 990–1015). Hoboken, NJ: John Wiley & Sons, Inc.

Taylor, R. M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *AGARD, Situational Awareness in Aerospace Operations* (pp. 90-28972–53).

Vortac, O. U., Edwards, M. B., Fuller, D. K., & Manning, C. A. (1995). Automation and Cognition in Air Traffic Control: An Empirical Investigation. *Applied Cognitive Psychology*, *7*, 631–651.

Wickens, C. D. (1996). Situation awareness: Impact of Automation and Display Technology. In *AGARD, Situational Awareness in Aerospace Operations* (Vol. 575, p. k2.1-k2.13). Brussels.

Wickens, C. D., Mccarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., Zheng, S., & Field, M. (2007). Attention-Situation Awareness (A-SA) Model of Pilot Error. In D. C. Foyle & B. L. Hooey (Eds.), *Human Performance Modeling in Aviation* (pp. 213–239). Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781420062984.ch9

Wogalter, M. S., & Laughery, K. R. (2012). Warnings and Hazard Communications. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics2* (4th ed., pp. 1191–1221). Hoboken, NJ: John Wiley & Sons, Inc.